Resolve, Capabilities, and Brinkmanship in Nuclear Crises

Brenton Kenkel[*] Peter Schram[†]

March 28, 2025

Abstract

Crises in the nuclear era are commonly framed as "brinkmanship," where actors compete by raising the background risk of a nuclear exchange until one side lacks the resolve to continue and backs down. But this framing may be too reductive: in practice, actors deploy a range of coercive capabilities that both alter the risk of escalation and shape political outcomes. How do these limited coercive capabilities shape outcomes in nuclear crises? We analyze the brinkmanship framework, finding broadly that more resolved actors will take greater escalation risks and perform better in conflict. We also analyze a "contests of capabilities" framework, showing that when a state's resolve also shapes its willingness to compete at lower levels, more resolved actors may engage in less risky or less decisive measures. We use a game-free methodology to study how the underlying military fundamentals affect crisis behavior in settings with autonomous escalation risk across a wide variety of bargaining games.

[*] Associate Professor, Department of Political Science, Vanderbilt University. brenton.kenkel@vanderbilt.edu

[†] Assistant Professor, Department of Political Science, Vanderbilt University. peter.schram@vanderbilt.edu

Nuclear weapons have redefined international relations (Brodie 1966; Schelling 1980; Jervis 1979; Bas and Coe 2012; Spaniel 2019). Consider coercive diplomacy in the nuclear era. Because a deliberate first strike using strategic nuclear weapons against a capable nuclear-armed adversary is effectively a suicidal act, except in matters of existential importance, states cannot wield these weapons as a credible direct threat (Schelling 1966). Instead, nuclear weapons shape coercive diplomacy through the possibility of their use. In crises in the nuclear era, states now take actions that carry the risk of misstep, miscalculation, or inadvertent escalation, all of which could lead to a catastrophic nuclear exchange (Brodie 1966; Schelling 1966, 1980; Powell 1989, 2015; Posen 2014). This new feature of great power competition has led scholars to classify crises and limited engagements among nuclear powers as exercises in "brinkmanship," where states use these underlying escalation risks to demonstrate their resolve to convince their opponents to stand down (Snyder 1965; Jervis 1976; Schelling 1980, 1966; Powell 1989, 1990). These brinkmanship contests resemble a game of "chicken" or of "rocking the canoe;" in them, adversaries seemingly stand eyeball to eyeball while escalating nuclear risks, hoping that their rivals are less willing to run such risks and blink first before disaster strikes.

The concept of brinkmanship has been incredibly useful in highlighting the new aspect of risk-taking in international politics in the nuclear era. However, much of what occurs in the nuclear era falls outside of this stylized framing. Consider the West's support for Ukraine in its war against Russia. By keeping Russia in a continued state of conflict, there exists the possibility that a human error or a faulty missile detection system results in an inadvertent escalation (Paul et al. 1990; Sagan 1994; Perrow 2011; Posen 2014). And, by running military supply chains through NATO countries, through mistake or malice, a Russian General may one day strike within a NATO state, thus raising the risk of a NATO-Russia war (Posen 2014). But despite these (even nuclear) escalation risks, the West's arming of Ukraine is not simply brinkmanship. The West arms Ukraine to protect the Ukrainian state and to strain the Russian state. This is not to say that such arming is not without nuclear risks, and these risks have undoubtedly influenced which weapons systems are transferred by the West. Instead, the West's strategy in Ukraine is shaped both by the political efficacy of supporting Ukraine and by the West's willingness to accepting the nuclear risks that stem from this support. And, the West's backing of Ukraine is not unique. Examples such as the Vietnam War, the Soviet Union's suppression of the Hungarian Revolution in 1956, and U.S. aid to the Afghan Mujahideen during the Soviet-Afghan War (1979–1989) all illustrate a recurring pattern: while nuclear risks play a part in shaping crisis behavior, these crises are much more than just leveraging nuclear risks to demonstrate resolve as described in classic brinkmanship. To understand state behavior and coercive politics in the nuclear era, a more general paradigm is needed.

We address this theoretical shortcoming. We begin by formally defining the brinkmanship theoretical framing. Within the brinkmanship framework, actors engage in a deterrence or bargaining game while taking actions that raise or lower the likelihood of a nuclear exchange—for example, by remaining in a conventional conflict that could, with some probability, escalate. This framing encompasses seminal models of nuclear deterrence like Nalebuff (1986), Powell (1989), and Powell (2015). Consistent with previous formal and informal characterizations of conflict behavior within brinkmanship, we find that *every* equilibrium of *every* brinkmanship game exhibits a similar pattern: if an actor is more resolved—that is, they privately value the asset in dispute more—then that actor takes on higher risks and attains higher expected rewards. To make this sweeping claim, we conduct a game-free analysis of the brinkmanship framing, along the along the lines of previous mechanism design research (Banks 1990; Fey and Ramsay 2011; Akçay et al. 2012; Spaniel 2020; Liu 2021).

We then present and analyze a more general theoretical framing: "contests of capabilities." In the contests of capabilities framing, actors have different conflict technologies at their disposal (like arming, engaging in limited war, conducting low-level operations, etc.) that, through their use, can generate nuclear escalation risks as well as latent costs. This differs from the brinkmanship framing where costs can only be generated through the manipulation of nuclear escalation risks. This new theoretical framing better represents the actions taken by states within crises, like arming allies or deploying conventional forces, that have their own costs outside of their effect on nuclear risks.[1] We find that these contests of capabilities exhibit very different behavioral patterns than brinkmanship contests: more resolved actors may be less likely to risk a nuclear exchange and may perform worse within the game.[2] This suggests that the patterns of behavior predicted in brinkmanship contests are, at best, special cases of how states will behave in crises in the nuclear era and, at worst, will predict the opposite of actual behavior.

Contests of capabilities predict new forms of crisis behavior because states are no longer limited to leveraging nuclear escalation risks as their only costly action. As intuition, consider what this game feature means for costly signaling, a common behavior within crisis bargaining and deterrence games. In models like Powell (2015)—which falls within our classification as a brinkmanship model—an actor may find it opportunistic to signal that they are high-resolve and willing to fight if challenged. Through the logic of costly signaling, they can only credibly do so by undertaking some action that results in them bearing some costs. Within Powell, the only costly action available is the generation of nuclear escalation risks; this means that, mechanically, in order to issue a costly signal of their resolve, more resolve actors must take on higher levels of nuclear risk. In contrast, within the contests of capabilities framework, actors can demonstrate their resolve by taking actions that generate costs outside of nuclear escalation risks, for example by investing in costly arming or by sending military aid to an ally, or implementing costly sanctions.[3] This new framework allows states

[1] As characterized here, a "brinkmanship crisis" is a special kind of "contest of capability" where there are no latent costs (i.e. no costs outside of nuclear risks).

[2] Note: this latter finding has not yet been proven in the main text, but it's basically been proven in Kenkel and Schram (2024).

[3] This is not to say that these actions would not potentially alter nuclear escalation risks; rather, how these actions alter nuclear escalation risks is a secondary concern. More theoretically, "burning cash" could be part of a contest of capability game—functioning as a costly signal without changing nuclear escalation risks—but not a brinkmanship game.

to construct new, non-nuclear costly signals, thus undermining the relationship described in brinkmanship games where more resolved actors take on greater risks. We demonstrate that the results within the brinkmanship framing can break down when states are able to undertake *any* costly action outside of manipulating nuclear risks, which, when considering the wide range of actions that states can take in crises, is an empirical certainty.

We make three primary contributions. First, this paper characterizes a new, generalized framework that better characterizes crisis behavior in the nuclear era. While the brinkmanship framework has been useful in highlighting new strategic tensions in the nuclear era, we show that it makes a critical oversimplification: it does not consider a state's ability to undertake any action that generates costs outside of the manipulation of nuclear risks. We remedy this by introducing the contests of capabilities framework, where a state's actions can generate both nuclear and non-nuclear costs. This new framework better matches the empirical reality that states face when engaging in crises in the nuclear era.

Second, this paper redefines how conflicts will play out in the nuclear era. Within the brinkmanship framework, the relationship between resolve and nuclear escalation risk largely follows what nuclear scholars have described in the past: more resolved actors are take on more nuclear risks in crises and perform better in these crises. However, we demonstrate that this relationship arises as an artifact of the overly simplistic cost structure within the brinkmanship framework. Once states can take actions that generate costs outside of nuclear escalation risks, more resolved states can engage in strategic behavior that is less risky. Why does this matter? Any game theoretic model is an exercise in simplification, but these simplifications should be made for tractability and should not prove decisive for shaping how players behave. If states can engage in *any* kind of costly behavior outside of manipulating nuclear risks—the existence of which is an empirical fact—then the patterns of behavior described in foundational nuclear deterrence research breaks. We demonstrate that the equilibrium behavior of resolved states is, generically, more complex than what was previously considered, but this behavior still follows predictable patterns that we characterize here.

Third, the results we present are extremely general, applying to *every* equilibrium of *every* game within the brinkmanship or contest of capabilities framing. A natural concern in analyzing an equilibrium to a game theory model is whether or not certain features of the game form are shaping strategic behavior in atypical ways. This is particularly important to the international relations setting, where there is no clear institutional framework that defines crisis bargaining behavior between states. We are able to avoid the pitfalls of analyzing a particular or esoteric game-form by utilizing the tools of mechanism design and conducting a game-form free analysis of both the brinkmanship and contests of capabilities framings. This analysis is novel, and the results presented here—through there generality—are also novel. While results like these exist in individual models like Powell (2015) (for the brinkmanship framing) and Schram (2024) (for the crisis of capabilities setting), we are the first to establish their ubiquity across each general class of games and to identify how and why they arise. And, while this kind of analysis has been applied to crisis bargaining models or to flexible-response crisis bargaining models in the past (Banks 1990; Fey and Ramsay 2011; Kenkel and Schram 2024), those analysis cannot speak to settings with a stochastic escalation risk.

We begin the more general analysis in Section 2. In Section 3, we offer a more straightforward grounding of how the assumptions of the brinkmanship and contests of capabilities framings produce distinct results.

1 Examples of Brinkmanship and Contest of Capabilities Games

Before we present our general results, we present examples of games that fall within the "brinkmanship" and the "contests of capabilities" frameworks. The brinkmanship game explored here is a simplified version of the Powell (2015) model.[4] The contest of capabilities game explored here is a slight modification of the brinkmanship model where D is able to generate costs outside of the nuclear escalation mechanism. Reasonable reviewers may take issue with some of the modeling choices in the games here. We chose to base these models on Powell (2015), a recent and well-cited model of nuclear brinkmanship, to give readers an understanding of how slight changes in that model can upend the equilibrium behavior described there. To summarize what follows, in the brinkmanship game, more-resolved actors risk nuclear escalation with greater likelihood. However, in a similar the contest of capabilities game where costs can be generated outside of the nuclear risk mechanism, this relationship no longer holds. After this analysis of these two games, in the following sections, we demonstrate that these equilibrium behaviors are not unique to the models explored here, but rather generic to a broad class of games.

1.1 Simple Brinkmanship Game

Consider a game within the brinkmanship framework, where a challenger (C) and a defender (D) are in a crisis over an asset. In this game, D is either resolved, placing a high value on the asset, or less resolved, placing a lower value on the asset. D's valuation of the asset is private, meaning that D knows their level of resolve, but C does not observe this. This game is a simplified version of the private-information game in Powell (2015), and the timing is as follows:

1. Nature selects D's private valuation $v_D \in \{\underline{v}_D, \bar{v}_D\}$. D is low-resolve ($\underline{v}_D$) with probability ζ and high-resolve ($\bar{v}_D$) with probability $1 - \zeta$.

2. D chooses a risk level $r \in \{\underline{r}, \overline{r}\}$, where $0 \leq \underline{r} < \overline{r} \leq 1$.

3. C may quit or challenge. If C quits, the game ends and C receives payoff $-k_C$ and D

[4] These simplifications—like assuming that D can select only low or high levels of risk rather than levels of risk from a continuum—are done to make the presentation of results cleaner. The equilibrium in Powell exhibits similar properties to the equilibrium discussed here.

receives payoff $v_D - k_D$, where $-k_C$ and $-k_D$ can be thought of as crisis costs. If C challenges, the game continues.

4. In response to C's challenge, D may quit or fight. If D quits, then C receives $v_C - k_C$ and D receives $-k_D$. If D fights, then C and D fight a war that could end conventionally or with a nuclear exchange. A nuclear war occurs with probability r and players receives $-n_i - k_i < 0$ (with $i \in \{C, D\}$) in this outcome. If a nuclear war does not occur, then D wins the asset with probability p, C wins the asset with probability $1 - p$, and both players incur the crisis costs $-k_i$.

To reduce the number of possible equilibria, make a series of assumptions. We assume that if D sets $r = \bar{r}$ and the game proceeds to stage 4 (i.e. C challenged in stage 3), then low-resolve D's prefer quitting to fighting and high-resolve D's prefer fighting to quitting. We also assume that if D sets $r = \underline{r}$ and the game proceeds to stage 4, then both types of D prefer fighting to quitting. Lastly, we assume that if D sets $r = \underline{r}$, then C prefers to challenge, even if doing so results in a fight.[5]

With these assumptions in place, a perfect Bayesian equilibrium exists. When C performs poorly in war, both types of D choose the high-risk level, leading C to consistently quit (Case A, formally described below). Alternatively, when C attains a high payoff from war, both types of D choose the low-risk level, C will challenge, and both types of D will fight (Case B). For a middle-range where C does well-but-not-too-well in war, two Cases exist. Sometimes both types of D's will prefer to set the low-risk level and always fight when challenged (Case C). Other times, a semi-separating equilibrium exists (Case D). In the semi-separating equilibrium, C will always challenge upon observing the low-risk level and will only sometimes challenge upon observing the right risk level. High-value D's will always choose the high-risk level and will fight when challenged. Low-value D's will mix between

[5] All of these assumptions are in place to reduce the number of equilibrium cases; without them, the relationship between resolve and nuclear risk would still hold in this model, and the results from our general analysis would still hold.

the high- and low-risk levels, and will only fight if challenged after selecting the low-risk level.

We provide a full equilibrium characterization in the Appendix; formally, on-the-path equilibrium play is the following. We define $\alpha = \frac{(1-\zeta)(\bar{r}n_C-(1-\bar{r})(1-p)v_C)}{\zeta v_C}$ and $\beta = \frac{\underline{v}_D-(1-\underline{r})p\underline{v}_D+\underline{r}n_D}{\underline{v}_D}$, which will be used below.

- Case A: If $(1-\zeta)\left((1-\bar{r})(1-p)v_C - \bar{r}n_C\right) + \zeta v_C \leq 0$, then both types of D set $r = \bar{r}$ and C always quits.
- Case B: If $(1-\bar{r})(1-p)v_C - \bar{r}n_C > 0$, then both types of D set $r = \underline{r}$, C always challenges, and both types of D fight.
- Case C: If $(1-\zeta)\left((1-\bar{r})(1-p)v_C - \bar{r}n_C\right) + \zeta v_C > 0$, $(1-\bar{r})(1-p)v_C - \bar{r}n_C \leq 0$, and $(1-\underline{r})p\bar{v}_D - \underline{r}n_D > (1-\beta)\bar{v}_D + \beta\left((1-\bar{r})p\bar{v}_D - \bar{r}n_D\right)$, then both types of D set $r = \underline{r}$, C always challenges, and both types of D fight.
- Case D: Otherwise, the equilibrium is semi-separating. Type $\bar{v}_D$ Ds will always set $r = \bar{r}$. Type $\underline{v}_D$ Ds will set $r = \bar{r}$ with probability α and will set $r = \underline{r}$ with probability $1-\alpha$. In response to $r = \underline{r}$, C will always challenge. In response to $r = \bar{r}$, C will challenge with probability β and not challenge with probability $1-\beta$. When C challenges after D sets $r = \underline{r}$, D will fight. If C challenges after D sets $r = \bar{r}$, type $\bar{v}_D$ will always fight and type $\underline{v}_D$ will always quit.

Consider the relationship between resolve and nuclear escalation risks. Across these equilibrium cases, a common pattern exists: if D has higher resolve, then, in equilibrium, D will have take on weakly higher levels of nuclear risks. For Cases A, B, and C, both types of D incur the same level of nuclear escalation risk. For Case D, the Appendix shows that that low-resolve Ds select a strategy where they face lower nuclear escalation risks than than high-resolve Ds. In this equilibrium, more resolved actors risk nuclear risks more often, which is consistent with past non-formal discussions of resolve. And, this result where more resolved actors risk nuclear war more often also appears in the Powell (2015) model, where actors have a richer action set, and in Nalebuff (1986), where actors engage in a continuous-time war-of-attrition and escalate nuclear risk by staying in the conflict for longer.

However, the model above makes a restrictive assumption: costs are only endogenously generated through the nuclear risk mechanism. In this simple model, like in Powell (2015), the defender can only incur costs through the manipulation of nuclear risks. Similarly, in Nalebuff (1986), actors only incurs costs by remaining in the crisis (which produces greater nuclear risks).[6] While the costs associated with nuclear escalation risks are important to consider—this modeling technology constitutes a meaningful departure from past formulations of conflict models—they are not the only way conflicts or crises are costly. For example, in these models, there is no external endogenous cost generated through arming, fighting conventionally, sanctions, or any other non-nuclear risk related mechanism. What happens, then, if these costs are included?

1.2 Simple Contests of Capabilities Game

Now consider a modified model where D can generate costs outside of the generation of nuclear risk within a war. In this new model, to generate the higher levels of nuclear escalation risk, D must pay a sunk cost. This modified model represents a simple depature from the technology used in the brinkmanship model above and in Nalebuff (1986) and Powell (2015).[7] As we will show, this new nature of costs can reshape equilibrium behavior.[8] The new model is as follows.

1. Nature selects D's private valuation $v_D \in \{\underline{v}_D, \bar{v}_D\}$. D is low-resolve ($\underline{v}_D$) with proba-

[6] Nalebuff (1986) describes dropping out as generating a "cost;" however, this "cost" is unavoidable when giving up the asset, making it a baseline payoff from not attaining the asset—ultimately, the cost from quitting in Nalebuff (1986) is analogous to the "quit" option and zero-payoff in in Powell (2015).

[7] What's unique here is that this cost is realized *regardless* of whether or not the conflict escalates. In the brinkmanship model, when war does not occur, there is no realized nuclear escalation risk.

[8] These sunk costs can have some empirical grounding. For example, if a domestic audience were to penalize a leader for escalating the risks of nuclear conflict, these costs could manifest in this way.

bility ζ and high-resolve ($\bar{v}_D$) with probability $1 - \zeta$.

2. D chooses a risk level $r \in \{\underline{r}, \bar{r}\}$, where $0 \leq \underline{r} < \bar{r} \leq 1$.

3. C may quit or challenge. If C quits, the game ends and C receives payoff $-k_C$ and D receives payoff $v_D - k_D - K(r)$, where $-k_C$ and $-k_D$ can be thought of as crisis costs and $K(r)$ are the costs to raising nuclear escalation risks. For ease, we assume $K(\bar{r}) > 0$ and $K(\underline{r}) = 0$.

4. In response to C's challenge, D may quit or fight. If D quits, then C receives $v_C - k_C$ and D receives $-k_D - K(r)$. If D fights, then C and D fight a war that could end conventionally or with a nuclear exchange. A nuclear war occurs with probability r, where C receives $-n_c - k_c < 0$ and D receives $-n_D - k_D - K(r)$. If a nuclear war does not occur, then D wins the asset with probability p, and C wins the asset with probability $1 - p$, D incurs costs $-k_D - K(r)$, C incurs costs $-k_C$.

We make a similar set of assumptions regarding D's stage 4 behavior as above and regarding C's preference for challenging when $r = \underline{r}$. What's new here is that we assume that low-resolve Ds and high-resolve Ds differ in their willingness to incur the sunk costs. We assume that high-resolve Ds prefer selecting the high nuclear escalation level (with high sunk costs) if this means that they attain the asset without a fight relative to selecting the low nuclear escalation level (with no sunk costs) and then fighting over the asset. We also assume that low-resolve types have the opposite preferences: they prefer selecting into the low nuclear escalation level and fighting over the asset to selecting the high nuclear escalation level and incurring the high sunk costs, even if it means that they attain the asset.

With these assumptions in place, a perfect Bayesian equilibrium exists. When C performs poorly in war, both types of D choose the low-risk level, and C consistently quits (Case E). Alternatively, when C attains a high payoff from war, both types of D choose the low-risk level, C will challenge, and both types of D will fight (Case F). Finally, for a middle-range where C does well-but-not-too-well in war, a fully separating equilibrium exists. In this equilibrium, high-value D's will always select the high risk level ($r = \bar{r}$) and C will not challenge, and low-value D's will always select the low risk level $r = \underline{r}$ and C will always challenge (Case G). We provide a full equilibrium characterization in the Appendix; formally, on-the-path equilibrium play is the following.

- Case E: If $0 < (1-\bar{r})(1-p)v_C - \bar{r}n_C$, then both types of D set $r = \underline{r}$, C always challenges, and both types of D fight.

- Case F: Otherwise, there is a separating equilibrium. Type $\bar{v}_D$ Ds will always set $r = \bar{r}$ and C will not challenge. And, type $\underline{v}_D$ Ds will always set $r = \underline{r}$, C will challenge, and D will fight.

Case F represents a new kind of behavior. In that equilibrium case, more resolved Ds are not running higher nuclear escalation risks—rather, by setting $r = \bar{r}$ and incurring costs $K(\bar{r})$, high resolve Ds have signalled that they are high-resolve, which results in C not challenging and the crisis ending without *any* nuclear escalation risks. This is in contrast to low-resolve Ds, who are not willing to incur the higher costs from setting $r = \bar{r}$, and will ultimately incur some escalation risks through fighting. This behavior is distinct from the classic thinking on brinkmanship, where more resolved actors will run higher nuclear risks. Rather, this equilibrium better resembles the signalling equilibrium in Schram (2024), which is similar to signaling equilibria in arming games, where more resolved types are more willing to engage in costly arming, which signals their type and deters challengers.

The models above are subject to critique; after all, they lack many important elements of crisis bargaining and deterrence theory, such as bargaining, cheap-talk signaling, or re-negotiating terms. And, while more sophisticated brinkmanship models (like Nalebuff (1986) and Powell (2015)) or contest of capabilities models (like Schram (2024)) exist, these models all (arguably) have their own shortcommings.[9] Furthermore, the kind of claims that we are

[9] For example, none of these models allow for bargaining.

interested in—when more resolved actors will (or will not) risk nuclear war with greater likelihood in crises—are sweeping and should exist beyond a single equilibrium of a single game form. To identify broad regularities in both classes of models, we employ the game-free methodology of Banks (1990) and Fey and Ramsay (2011), identifying properties that arise from foundational requirements of equilibrium rather than idiosyncratic features of any given game tree. We find, in fact, that the intuitive relationship between resolve and war risk in the brinkmanship game above holds in every equilibrium of every game within the brinkmanship framework. And, we also find that the counterintuitive relationship between resolve and war risk in the contests of capabilities game can emerge in a wide range of games in the contests of capabilities setting.

2 Model Framework

To model contests of capabilities, we extend the formal definition of flexible-response crisis bargaining games developed by Kenkel and Schram (2024). A contest of capabilities is a game between a Challenger C and a Defender D, who are in a crisis over a prize whose value is normalized to 1.[10] The two sides bargain over the division of the prize. In the course of bargaining, each side may take costly low-level actions to influence the outcome of negotiations. What's novel here—and not considered in Kenkel and Schram (2024)—is that D's low-level action may generate a risk of accidental escalation to full-scale war.[11] If C and D do not agree on a negotiated settlement, then war occurs for certain. D has private information about its war payoff, which creates a friction that may result in bargaining failure in equilibrium.

[10] In section 5 below, we consider an alternative formulation where D's valuation of the prize is private information that varies across private types.

[11] The flexible-response crisis bargaining games studied by Kenkel and Schram (2024) are the special case of contests of capabilities in which this risk is identically zero.

2.1 Contests of Capabilities

The timing of a contest of capabilities is as follows. At the start of the game, Nature draws D's type θ from a commonly known distribution whose CDF is F_θ and whose support is $\Theta \subseteq \mathbb{R}$. Note that we first characterize type in terms of costs; see section 5 for a treatment of resolve as valuation. Only D observes Nature's choice. The two players select bargaining strategies $b_C \in \mathcal{B}_C$ and $b_D \in \mathcal{B}_D$. Our analysis is agnostic as to the shape of the bargaining protocol. Depending on the particular game form, these strategies might be simple proposals and responses, as in an ultimatum game, or more complex plans of offers and counteroffers. Alongside these baseline bargaining strategies, each player may take a costly action that directly shifts the payoffs from negotiations—and, in D's case, may generate a heightened risk of accidental escalation to war. We call C's action a "transgression" $t \in \mathcal{T} \subseteq \mathbb{R}_+$ and D's "hassling" $h \in \mathcal{H} \subseteq \mathbb{R}_+$.

The players' bargaining strategies determine whether the negotiation succeeds or fails. For any given game form g, there is a function $\pi^g : \mathcal{B}_C \times \mathcal{B}_D \to [0,1]$ that describes the probability of agreement (i.e., neither player deliberately choosing war) given the players' bargaining strategies. In case of agreement, C pays a cost $K_C(t) \geq 0$ (increasing in t) for its transgressions, and D pays a potentially type-dependent cost $K_D(h, \theta) \geq 0$ (increasing in h) for its hassling.[12] In this case, the probability of accidental escalation to war is a strictly increasing function of D's hassling, denoted $R(h) \in [0,1]$. We treat the costs of transgression and hassling, as well as the risk of accidental escalation, as underlying primitive features of the strategic interaction—not as features of a single particular game form g. If there is no accidental escalation to war, the players receive $V_C^g(t, h, b_C, b_D)$ and $V_D^g(t, h, b_C, b_D)$ respectively. Unlike the cost and risk functions, these are specific to a game form.

[12] Because we conceptualize accidental escalation as arising from the coercive instrument itself, we assume these costs are paid even if accidental escalation occurs. For example, if a conventional war ends in an accidental nuclear exchange, the conventional war still carries costs (as was similarly formalized in Powell (2015)). Consequently, both players would prefer deliberate war—avoiding the costs of the low-level actions—over an agreement with a near-certain chance of accidental escalation.

Like earlier game-free analyses of crisis bargaining (Fey and Ramsay 2011; Fey and Kenkel 2021; Kenkel and Schram 2024), we assume that both states have the option to unilaterally force a conflict. Formally, this amounts to assuming there exists $b_C^{\text{war}} \in \mathcal{B}_C$ such that $\pi^g(b_C^{\text{war}}, b_D) = 0$ for all $b_D \in \mathcal{B}_D$, as well as an analogous $b_D^{\text{war}} \in \mathcal{B}_D$. Reflecting the anarchic nature of international politics, this assumption ensures that neither state can be forced to accept a settlement that would leave it worse off than fighting.

The players' baseline war payoffs are solely a function of D's type, not their bargaining actions or low-level responses (see section 5 for an alternate treatment). We order the Defender's type space so that higher types are more resolved, i.e., the Defender's baseline war payoff function $W_D(\theta)$ is a strictly increasing function of θ. Meanwhile, the Challenger's baseline war payoff $W_C(\theta)$ is a non-increasing function of θ. If war occurs deliberately due to bargaining failure, then the players receive $W_C(\theta)$ and $W_D(\theta)$. If war occurs accidentally due to hassling-induced escalation, then they receive $W_C(\theta) - K_C(t)$ and $W_D(\theta) - K_C(h, \theta)$. Notice that transgressions and hassling are the only bargaining actions that may affect war payoffs, and they do so only for accidental escalation and only via the cost functions.

The Challenger's expected utility, given the bargaining strategies and the Defender's type, is given by the function

$$u_C^g(t, h, b_C, b_D \mid \theta) = \underbrace{\pi^g(b_C, b_D)}_{\text{agreement}} \left[\overbrace{[1 - R(h)] V_C^g(t, h, b_C, b_D)}^{\text{no escalation}} + \overbrace{R(h) W_C(\theta)}^{\text{accidental escalation}} - K_C(t) \right] + \underbrace{[1 - \pi^g(b_C, b_D)]}_{\text{disagreement}} W_C(\theta).$$

Similarly, the Defender's expected utility function is

$$u_D^g(t, h, b_C, b_D \mid \theta) = \pi^g(b_C, b_D) \left[[1 - R(h)] V_D^g(t, h, b_C, b_D) + R(h) W_D(\theta) - K_D(h, \theta) \right] + [1 - \pi^g(b_C, b_D)] W_D(\theta).$$

To close the definition of contests of capabilities, we place some additional assumptions on the model primitives. First, we assume that either player may refrain from low-level responses at no cost: $0 \in \mathcal{H} \cap \mathcal{T}$, $K_C(0) = 0$, and $K_D(0, \theta) = 0$ for all $\theta \in \Theta$. Second, we assume there is no risk of accidental escalation in the absence of hassling: $R(0) = 0$. Third, without loss of generality, we let $W_D(\theta) = \theta$ in the remainder of the analysis.

2.2 Game-Free Analysis

Our goal is to characterize patterns in the equilibria of contests of capabilities that hold across all game forms with the same underlying primitives, rather than being specific to a particular bargaining protocol (e.g., ultimatum game, alternating offers, etc.). Table 1 divides the model components into the primitive components and those that are specific to a particular game form. We will draw conclusions about the equilibrium outcomes of contests of capabilities solely as a function of the primitive components listed in the left-hand column. To this end, we adopt the mechanism design methodology of prior game-free analyses of crisis bargaining (Banks 1990; Fey and Ramsay 2009, 2011; Fey and Kenkel 2021; Liu et al. 2021; Kenkel and Schram 2024).

Underlying primitives	**Specific to game form**
D's type space: Θ	Bargaining actions: $\mathcal{B}_C$, $\mathcal{B}_D$
War payoffs: $W_C(\cdot)$, $W_D(\cdot)$	Actions → bargaining success: $\pi^g(\cdot)$
Low-level responses available: $\mathcal{T}$, $\mathcal{H}$	Actions → prize division: $V_C^g(\cdot)$, $V_D^g(\cdot)$
Cost functions: $K_C(\cdot)$, $K_D(\cdot)$	
Escalation risk: $R(\cdot)$	

Table 1: Classification of model components for a contest of capabilities.

As in similar analyses of models with one-sided incomplete information (e.g., Banks 1990; Fey and Kenkel 2021; Kenkel and Schram 2024), we characterize equilibrium outcomes for the player with private information, namely the Defender. Let $(t^*, h^*(\theta), b_C^*, b_D^*(\theta))$ be an equilibrium of a game form g, where D's strategies are written as functions of θ as different types may take different actions. We summarize the equilibrium via three functions of D's type. The first is the equilibrium probability of agreement for each Defender type:

$$\pi(\theta) = \pi^g(b_C^*, b_D^*(\theta)).$$

The second is the equilibrium amount of hassling, conditional on an agreement, for each Defender type:

$$h(\theta) = h^*(\theta).$$

The third is the equilibrium division of spoils going to the Defender, conditional on an agreement and no accidental escalation to war, for each type:

$$V_D(\theta) = V_D^g(t^*, h^*(\theta), b_C^*, b_D^*(\theta)).$$

We refer to these three functions $(\pi(\cdot), h(\cdot), V_D(\cdot))$ jointly as a direct mechanism. Rather than work with the complex set of all equilibria of all game forms, we will work with the set of direct mechanisms for contests of capabilities.

Given a direct mechanism, we can calculate the expected utility to each type of Defender as follows:

$$U_D(\theta) = \pi(\theta)\left[(1 - R(h(\theta)))V_D(\theta) + R(h(\theta))\theta - K_D(h(\theta), \theta)\right] + (1 - \pi(\theta))\theta.$$

Equally importantly, the direct mechanism gives us all we need to know to determine the payoff one Defender type would receive by deviating to another type's bargaining strategy. Consider a Defender whose true type is θ, but who mimics the bargaining strategy of another type θ'. This type receives the same lottery over agreement versus disagreement (probability $\pi(\theta')$ of agreement) and the same risk of accidental war in case of agreement (probability $R(h(\theta'))$) as the type it is mimicking. Additionally, if there is agreement and no accidental war, the mimicking type receives the same bargaining spoils, $V_D(\theta')$. But there are two key differences between the mimic's payoff and $U_D(\theta')$. First, in case of war (whether accidental or deliberate), the mimic receives its true private value θ—it does not become stronger or weaker on the battlefield just by adopting the bargaining strategy of a different type. Second, in case of agreement, the type-dependent component of the mimic's hassling cost reflects its true type; i.e., the mimic pays $K_D(h(\theta'), \theta)$. Altogether, then, the expected utility to type θ for adopting the bargaining strategy of type θ' is

$$\Phi_D(\theta' \mid \theta) = \pi(\theta') \left[(1 - R(h(\theta')))V_D(\theta') + R(h(\theta'))\theta - K_D(h(\theta'), \theta)\right] + (1 - \pi(\theta'))\theta.$$

A direct mechanism is incentive compatible if no Defender type would strictly benefit from mimicking the bargaining strategy of a different type. Formally, the incentive compatibility condition is

$$U_D(\theta) \geq \Phi_D(\theta' \mid \theta) \qquad \text{for all } \theta, \theta' \in \Theta. \tag{IC}$$

The incentive compatibility condition is closely related to the equilibrium requirements of Bayesian games. Our game-free analysis depends critically on the revelation principle articulated by Myerson (1979): for every Bayesian Nash equilibrium of a Bayesian game, there exists a payoff-equivalent direct mechanism that is incentive compatible. Therefore, if some claim holds for all incentive compatible direct mechanisms for contests of capabilities, then the same claim is true for all equilibria of such contests. By analyzing the set of incentive compatible direct mechanisms, we can derive necessary conditions for equilibrium behavior without having to solve any specific game form.

In line with Fey and Ramsay (2011) and the subsequent mechanism design literature, we also impose a voluntary agreements condition—what economists would call a participation or individual rationality constraint—on the set of direct mechanisms we consider. Voluntary agreements holds when no Defender type is worse off than it would be from deliberately fighting:

$$\pi(\theta)\left[(1-R(h(\theta)))V_D(\theta)+R(h(\theta))\theta-K_D(h(\theta),\theta)\right]\geq\pi(\theta)\theta \qquad \text{for all } \theta\in\Theta. \tag{VA}$$

Formally, voluntary agreements is a consequence of our assumption that both players have a bargaining action available that guarantees war. Substantively, this condition reflects the anarchic state of international politics, in which all agreements must be self-enforcing. Voluntary agreements hold trivially for any type that deliberately chooses war for certain, i.e., for which $\pi(\theta)=0$. Additionally, if we have $\pi(\theta)=0$ for at least one Defender type, then incentive compatibility implies voluntary agreements.

3 Brinkmanship Contests

We conceptualize a contest of nerves as a special case of a contest of capabilities, in which all Defender types have the same access to an instrument that generates exogenous escalation risks. In brinkmanship, Defender types may vary in their resolve—their expected value of fighting, and thus their willingness to risk war in order to receive a particular settlement at the bargaining table—but no type has an advantage over any other at pulling the levers that generate risk. In its simplest form, the risk of accidental war can be thought of as a pure brinkmanship measure, akin to rocking the boat in Schelling (1966, 90-91). This pure brinkmanship dynamic is formalized in models like Nalebuff (1986), Powell (1988), and Powell (1990).[13] Additionally, the brinkmanship framing can also describe settings where lower-levels actions may be inefficient, so long that these inefficiencies are uncorrelated with private type. Also, as discussed in section 5, this framework closely relates to the model in Powell (2015).

[13] Our framing differs from these models in our treatment of "resolve," which only depends on D's escalated war payoff. For example, in Powell (1988), resolve is a function of war payoffs, the payoffs from prevailing in the crisis, and the payoffs from conceding in the crisis; we choose a simpler treatment of resolve that does not rely on factors like the payoffs from dropping out of a crisis, which, in the bargaining setting, may be endogenous to the Defender's war payoff.

Formally, we define a contest of nerves as one in which the cost of each feasible low-level choice is constant across Defender types. In a contest of nerves, there exists a non-decreasing function $\kappa_D : \mathcal{H} \to \mathbb{R}_+$ such that

$$K_D(h, \theta) = \kappa_D(h) \qquad \text{for all } h \in \mathcal{H} \text{ and } \theta \in \Theta.$$

This includes as a special case "pure brinkmanship" scenarios, in which all types can generate accidental war risk costlessly: $\kappa_D(h) = 0$ for all $h \in \mathcal{H}$ (like in Nalebuff (1986) and Powell (1988). When the Defender's private information is about its war payoff, the distinction between the pure brinkmanship model and the contest of nerves turns out to be essentially immaterial.[14]

Brinkmanship contests exhibit essentially the same patterns of behavior as in ordinary crisis bargaining games, where there are no low-level policy alternatives between peaceful settlement and all-out war. When accounting for both of the possible paths to conflict—deliberate war or accidental escalation—more resolved types of the Defender are more likely to end up at war in equilibrium. Additionally, more resolved types have higher equilibrium payoffs. The following proposition states these claims formally as properties of incentive compatible direct mechanisms for brinkmanship.[15]

Proposition 1. *In any equilibrium of a contest of nerves, the total probability of war and the Defender's equilibrium utility weakly increase with the Defender's resolve: if $\theta' < \theta''$, then $\pi(\theta')[1 - R(h(\theta'))] \geq \pi(\theta'')[1 - R(h(\theta''))]$ and $U(\theta') \leq U(\theta'')$.*

Proposition 1 holds because brinkmanship contests are, in fact, equivalent to ordinary crisis bargaining games at a deep level. Because the cost of the low-level option does not differ

[14] The same is not necessarily true when the Defender's private information concerns its prize valuation; see section 5 below.

[15] All proofs appear in Appendix C.

across types, any type that mimics the bargaining strategy of θ' receives exactly the same payoff in case of a peaceful outcome, namely $V_D(\theta') - \kappa_D(h(\theta'))$. This equivalence of settlement payoffs across types means that the monotonicity results from Banks (1990) apply to brinkmanship contests, so higher types of the Defender are more likely to go to war and have greater equilibrium expected utilities. By contrast, in the more general class of contests of capabilities that we study below, different Defender types might yield different payoffs from the same bargaining strategy, even conditional on the interaction ending peacefully. That is because the costs of the low-level policy may differ across types, leading to different overall settlement values under which the Banks (1990) results no longer apply (see Kenkel and Schram 2024).

When full-scale war is sufficiently destructive, such as a nuclear exchange would be, it is plausible to suppose neither side would ever deliberately initiate a conflict (Brodie 1966; Schelling 1966; Powell 1990). In this case, we can obtain even stronger results about the relationship between the Defender's private resolve and equilibrium choices. Specifically, more resolved types engage in more brinkmanship and receive more favorable settlements at the bargaining table when the game ends peacefully.

Corollary 1. *In any equilibrium of a contest of nerves, if war never occurs deliberately ($\pi(\theta) = 1$ for all θ), then the probability of accidental war and the Defender's settlement value weakly increase with the Defender's resolve: if $\theta' < \theta''$, then $R(h(\theta')) \leq R(h(\theta''))$ and $V_D(\theta') \leq V_D(\theta'')$.*

Altogether, in a contest of nerves in which no Defender type has a particular advantage or disadvantage at generating accidental war risk, outcomes are determined by resolve in a predictable way. Greater resolve implies a greater total risk of war, including a greater risk of accidental war when neither player would ever deliberately opt into conflict. But as we show below, this stark pattern does not necessarily hold in more general contests of capabilities, where we consider accidental war risk a byproduct of low-level policy responses whose cost or effectiveness is related to the Defender's resolve.

4 Contests of Capabilities

In the contest of capabilities framework, the Defender's private willingness to go to war is related to their ability or willingness to use limited instruments. This creates a potential new tradeoff that affects the Defender's equilibrium choices. Types may vary in their preferences for lower-level conflict not only due to the risk of full-scale war that these options generate, but also because of differences in their direct costs for using these instruments. Consequently, even when full-scale war never occurs on purpose, it is no longer certain that more resolved types take on a higher risk of accidental conflict.

Throughout this section, we restrict attention to direct mechanisms in which each Defender type is either certain to agree or certain to go to full-scale war: $\pi(\theta) \in \{0, 1\}$ for all $\theta \in \Theta$. This set of mechanisms corresponds to pure-strategy equilibria of contests of capabilities in which Nature's only moves are the initial assignment of the Defender's type and the risk of war $R(h)$ generated by the Defender's low-level policy choice (see Fey and Kenkel 2021). Given such a mechanism, we can partition the type space into those that reach agreement (with possible risk of accidental conflict) and those that deliberately fight a war: $\Theta = \Theta_1 \cup \Theta_0$, where $\Theta_1 = \{\theta \in \Theta \mid \pi(\theta) = 1\}$ and $\Theta_0 = \{\theta \in \Theta \mid \pi(\theta) = 0\}$. We relax this assumption, allowing for equilibria in which some Defender types mix, in subsection 4.3 below.

4.1 Probability of Accidental War

Unlike in the special case of the contest of nerves analyzed above, the probability of accidental war need not increase with the Defender's resolve in a contest of capabilities. To see why, consider the tradeoff between low and high hassling, and how it varies with the Defender's resolve. First, there will be a difference in the Defender's settlement value if accidental war does not occur.[16] The Defender's resolve is immaterial to the value of this difference. Second, higher hassling generates a greater risk of accidental conflict. This is the effect at the root of Corollary 1 above, leading more resolved types to be more tolerant of greater hassling. But now there is a third component to the tradeoff: the marginal cost of the higher value of hassling may differ with the Defender's resolve. In practical terms, some Defenders may be more or less willing to conduct low-level competition depending on their high-level resolve; for example, if a Defender is more hawkish and willing to risk a nuclear exchange, then this Defender could plausibly also be more willing to absorb the costs from a more expansive conventional conflict. Naturally, we cannot characterize the relationship between resolve and equilibrium behavior without accounting for these marginal costs—and how they compare to the increase in the risk of accidental war.

If the marginal cost of hassling decreases with the Defender's resolve, then it is straightforward to see that more resolved types will run a greater risk of accidental war. In this case, compared to a less resolved Defender, a more resolved type gets the same benefit in case the agreement holds, is better off in case accidental war occurs, and pays less to go from low hassling to high hassling.

If instead more resolved Defender types face higher marginal costs of hassling (e.g., because investments in capabilities for total war crowd out the resources for lower-level instruments), then the tradeoff is harder to resolve. More resolved types are still better able to handle the risk of accidental war, but now they must pay more to demonstrate this resolve through low-level conflict. Ultimately, equilibrium behavior here comes down to the relative magnitude of (a) the effect of Defender resolve on the marginal cost of hassling and (b) the effect of hassling on the risk of accidental war. If the increase in the low-level instrument does not generate much risk, then we see the opposite pattern from the war of nerves, with more resolved Defenders investing less in low-level conflict and thus facing lower odds of accidental war.

[16] Intuitively, one might expect more hassling to yield more favorable terms. In fact, additional assumptions are required to guarantee this is the case. See Kenkel and Schram (2024).

The following proposition states sufficient conditions for the equilibrium probability of accidental war to increase or decrease with the Defender's resolve among all types that reach an agreement. The left-hand side of the equations in the proposition is, in essence, the effect of resolve on the marginal cost of greater hassling. The right-hand side is the effect of greater hassling on the risk of accidental conflict, weighted by the difference in war payoff between the more resolved and the less resolved type.

Proposition 2. *Consider an equilibrium of a contest of capabilities.*

(a) The probability of accidental war weakly increases with the Defender's resolve if

$$[K_D(h'', \theta'') - K_D(h', \theta'')] - [K_D(h'', \theta') - K_D(h', \theta')] < [R(h'') - R(h')](\theta'' - \theta') \quad (1)$$

for all $h', h'' \in h(\Theta_1)$ and $\theta', \theta'' \in \Theta_1$ such that $h' < h''$ and $\theta' < \theta''$.

(b) The probability of accidental war weakly decreases with the Defender's resolve if

$$[K_D(h'', \theta'') - K_D(h', \theta'')] - [K_D(h'', \theta') - K_D(h', \theta')] > [R(h'') - R(h')](\theta'' - \theta') \quad (2)$$

for all $h', h'' \in h(\Theta_1)$ and $\theta', \theta'' \in \Theta_1$ such that $h' < h''$ and $\theta' < \theta''$.

Equation 1 and Equation 2 are closely related to the single-crossing conditions that often arise in mechanism design and related economic settings (Milgrom and Shannon 1994; Ashworth and Bueno de Mesquita 2006). If K_D has global decreasing differences—i.e., the increase in cost between any two levels of hassling is always smaller for more resolved types—then Equation 1 must hold, and the probability of accidental war must behave as it does in a contest of nerves. On the other hand, if K_D has global increasing differences, then accidental war risks may still increase with resolve; a slim increase in marginal costs with θ is not enough to make Equation 2 hold.[17]

The key takeaway from Proposition 2 is that the direction of the relationship between resolve and accidental war depends on how resolve affects the marginal cost of hassling versus how hassling affects the risk. Can we say anything stronger about the shape of this relationship, such as how quickly the level of low-level activity varies with the Defender's resolve? Our next result, Proposition 3, answers this question in the negative. As long we can satisfy the conditions on marginal effects set out in the previous proposition, we can design a game form to rationalize virtually any pattern of hassling.

To obtain the following result, we must impose slightly stronger technical conditions than in the baseline analysis. We assume a continuous type space, $\Theta = [\underline{\theta}, \overline{\theta}]$, and set of feasible low-level actions, $\mathcal{H} = [0, \overline{h}]$. We also assume that R is continuously differentiable and that K_D is twice differentiable. Together we refer to these as the differentiability assumptions.

Proposition 3. *Suppose the differentiability assumptions hold.*

(a) Let h^* be any absolutely continuous, weakly increasing function that satisfies

$$\frac{\partial^2 K_D(h,\theta)}{\partial h \partial \theta} \leq R'(h) \qquad \textit{for all } h \in h^*(\Theta),\ \theta \in \Theta.$$

There is an incentive compatible direct mechanism in which $\pi(\theta) = 1$ and $h(\theta) = h^*(\theta)$ for all $\theta \in \Theta$.

(b) Let h^{**} be any absolutely continuous, weakly decreasing function that satisfies

$$\frac{\partial^2 K_D(h,\theta)}{\partial h \partial \theta} \geq R'(h) \qquad \textit{for all } h \in h^{**}(\Theta), \theta \in \Theta.$$

[17] K_D need not have global increasing or decreasing differences. For example, consider the type space $\Theta = [0,1]$, hassling space $h = \{0,1,2\}$, and cost function $K_D(0,\theta) = 0$, $K_D(1,\theta) = a + b\theta$ (where $0 < a < 1$ and $0 < b < 1 - a$), $K_D(2,\theta) = 1$. This function has increasing differences on $\{0,1\}$, decreasing differences on $\{1,2\}$, and constant differences on $\{0,2\}$. Consequently, if the effect of hassling on the risk of accidental war is strong enough, there may simultaneously exist IC mechanisms with increasing hassling (on $\{1,2\}$) and decreasing hassling (on $\{0,1\}$).

There is an incentive compatible direct mechanism in which $\pi(\theta) = 1$ and $h(\theta) = h^{**}(\theta)$ for all $\theta \in \Theta$.

Though the formal statement is technical, this is a remarkable result. Proposition 3 tells us that the primitive features of the strategic environment—i.e., the players' types, the low-level actions available, the costs of those actions, and the risk of accidental war they generate—only determine whether the equilibrium degree of hassling (and accidental war) increases or decreases with the Defender's resolve. The precise magnitude of the increase or decrease may be quite specific to a particular bargaining game form. For example, suppose that K_D has global decreasing differences (more resolved types face lower marginal costs of hassling), so that $\frac{\partial^2 K_D(h,\theta)}{\partial h \partial \theta} < 0$ for all h and θ. Then Proposition 3(a) implies that *every* increasing function $h^*(\theta)$ (subject to a continuity restriction) can be supported as equilibrium behavior. Whether there is a gradual increase, a sudden large spike, or no change at all depends only on the contingent features of the bargaining game, not on the underlying military fundamentals.

4.2 Occurrence of Deliberate War

We now consider how the choice of deliberate war varies with the Defender's resolve. By definition, the payoff from war is greater for more resolved types, increasing their incentive to opt for war. In the absence of low-level alternatives, this incentive results in a monotone increasing relationship between resolve and deliberate war (Banks 1990). However, in a more general setting, the effect of resolve on deliberate war depends on its relationship with the cost of low-level policy options. In particular, if more resolved types can also use low-level conflict more cheaply or effectively, then they may opt for limited conflict instead of war (Kenkel and Schram 2024).

While the relationship between resolve and hassling costs is important, the risk of accidental war from limited conflict also plays a role in the decision to fight war deliberately. If a limited policy instrument generates only an infinitesimal risk of accidental war, then the types with the most incentive to opt for it will be whichever ones have the lowest costs for that level of hassling. At the other extreme, if limited conflict carries an enormous risk of escalation, then resolve rather than costs will be the determining factor in the preference of each Defender type.

The following proposition summarizes the relationship between Defender resolve and the occurrence of deliberate war in the equilibria of contests of capabilities. If more resolved types pay higher (absolute) costs for low-level policy options, then less-resolved types settle and more-resolved types fight. However, when greater resolve also entails greater willingness or capability to hassle, then we must compare the magnitude of the cost effect to the degree of accidental war risk. If resolve only slightly reduces the hassling costs, or if the risk of accidental escalation is low, then we still have less-resolved types agreeing and more-resolved types fighting. But we yield the opposite pattern if hassling costs plunge quickly with θ or if escalation risks are high.

Proposition 4. *Consider an equilibrium of a contest of capabilities. Let Θ_1 denote the set of types that reach an agreement in equilibrium: $\Theta_1 \equiv \{\theta \in \Theta \mid \pi(\theta) = 1\}$.*

(a) Less-resolved Defender types reach agreement and more-resolved types deliberately choose war (i.e., $\operatorname{clos} \Theta_1 = \{\theta \in \Theta \mid \theta \leq \hat{\theta}\}$ for some $\hat{\theta}$) if

$$\frac{K_D(h', \theta') - K_D(h', \theta'')}{\theta'' - \theta'} < 1 - R(h') \tag{3}$$

for all $h' \in h(\Theta_1)$ and all $\theta', \theta'' \in \Theta$ such that $\theta' < \theta''$.

(b) Less-resolved Defender types deliberately choose war and more-resolved types reach agreement (i.e., $\operatorname{clos} \Theta_1 = \{\theta \in \Theta \mid \theta \geq \hat{\theta}\}$ for some $\hat{\theta}$) if

$$\frac{K_D(h', \theta') - K_D(h', \theta'')}{\theta'' - \theta'} > 1 - R(h') \tag{4}$$

for all $h' \in h(\Theta_1)$ and all $\theta', \theta'' \in \Theta$ such that $\theta' < \theta''$.

We briefly note the technical similarities and differences between Proposition 2 (conditions for the probability of accidental war to be monotone in the Defender's resolve) and Proposition 4. Both results work with differences in the cost function, $K_D(\cdot, \theta'') - K_D(\cdot, \theta')$, as well as in war payoffs, $\theta'' - \theta'$. Additionally, the risk of accidental war plays a role in both results. The probability of accidental war characterized in Proposition 2 is ultimately determined by a second-order comparison: the effect of Defender type on the *marginal* cost of hassling, compared to the marginal effect of hassling on escalation risk. By contrast, the occurrence of deliberate war characterized here in Proposition 4 depends more on first-order comparisons: the effect of Defender type on the *absolute* cost of hassling, versus the absolute level of escalation risk.

4.3 Probabilistic Deliberate War

Our baseline analysis concerns equilibria in which every Defender type either reaches agreement for certain or fights a deliberate war for certain. In such equilibria, the Defender can generate a limited risk of war only through choosing a corresponding level of hassling, not through mixed strategies. We now relax this restriction to consider equilibria in which we may have $\pi(\theta) \in (0, 1)$ for some (or all) types of the Defender.

When we allow for a probabilistic occurrence of deliberate war, we find exceptions to some of the patterns characterized in our baseline analysis. For example, consider an environment in which the effect of resolve on hassling cost is negative (Equation 3 holds), and this effect becomes larger in magnitude at higher degrees of hassling (Equation 1 holds). Under these conditions, Proposition 2 would lead us to conclude that the probability of accidental war increases with the Defender's resolve, and we would infer from Proposition 4 that the same is true for the occurrence of deliberate war. However, these patterns are specific to equilibria in which the probability of deliberate war is exactly 0 or 1 for each Defender type. Proposition 5 below shows that the probability of accidental escalation may decrease with Defender resolve when the probability of deliberate war is locally increasing. Alternatively, if the probability of accidental escalation increases quickly enough, then the probability of deliberate war may decrease with Defender resolve.

For ease of characterization, we impose linearity assumptions that are even stronger than the differentiability assumptions of Proposition 3 above. We assume $\Theta = [\underline{\theta}, \overline{\theta}]$ and $\mathcal{H} = [0, 1]$.[18] Additionally, we assume the cost function and risk function are linear in hassling: there exist a function $k : \Theta \to \mathbb{R}_{++}$ and a constant $r > 0$ such that $K_D(h, \theta) = k(\theta)h$ and $R(h) = rh$ for all $\theta \in \Theta$ and $h \in \mathcal{H}$. We also assume k is differentiable, and we denote $\underline{k} \equiv \min k'(\Theta)$ and $\overline{k} \equiv \max k'(\Theta)$.

Proposition 5. *Suppose the linearity assumptions hold and $r - 1 < \underline{k} \leq \overline{k} < r$.*

(a) Equation 1 and Equation 3 hold.

(b) Let π^* be any absolutely continuous, weakly decreasing function such that $\pi^*(\theta) > 0$ for all $\theta \in \Theta$, and let h^* be any absolutely continuous function that satisfies

$$\frac{dh^*(\theta)}{d\theta} \geq \frac{\frac{1}{r-\underline{k}} - h^*(\theta)}{\pi^*(\theta)} \cdot \frac{d\pi^*(\theta)}{d\theta}$$

for all $\theta \in \Theta$ at which h^* and π^* are differentiable. There is an incentive compatible direct mechanism in which $\pi(\theta) = \pi^*(\theta)$ and $h(\theta) = h^*(\theta)$ for all $\theta \in \Theta$.

(c) Let π^{**} be any absolutely continuous, weakly increasing function such that $\pi^{**}(\theta) > 0$ for all $\theta \in \Theta$, and let h^{**} be any absolutely continuous function that satisfies

$$\frac{dh^{**}(\theta)}{d\theta} \geq \frac{\frac{1}{r-\overline{k}} - h^{**}(\theta)}{\pi^{**}(\theta)} \cdot \frac{d\pi^{**}(\theta)}{d\theta}$$

[18] Because the cost function and war payoff space are arbitary in magnitude, the normalization $\overline{h} = 1$ is without loss of generality.

for all $\theta \in \Theta$ at which h^{**} and π^{**} are differentiable. There is an incentive compatible direct mechanism in which $\pi(\theta) = \pi^{**}(\theta)$ and $h(\theta) = h^{**}(\theta)$ for all $\theta \in \Theta$.

Though it is important to understand these baseline exceptions to the patterns characterized in our main analysis, the takeaway here should not be that anything can happen. In substantively important contexts such as nuclear war where it is implausible that the Defender would ever deliberately opt for war, we cannot obtain these exceptions to Proposition 2. Additionally, the result here depends on the cost effect being in a tight range where it is small in magnitude. A larger negative or positive relationship between resolve and hassling cost would tighten the set of patterns that can be sustained in equilibrium.

5 Resolve as Prize Value

The analysis above considers a class of models that treat the Defender's resolve as their war payoff. Here, we consider an alternate formulation, more in line with recent nuclear brinkmanship models (e.g., Powell 2015; Schram 2024), in which the Defender's private resolve is represented by their value for the object at stake in the crisis. We are able to show that this alternate formulation can share similarities with the contest of capabilities framework, but, with some new model primitives (i.e. a option to quit and zero out), can also introduce new "outbidding" behavior.

To motivate the analysis, consider a simple model of brinkmanship based on Powell (2015).[19] C and D are in a crisis over a prize worth $\beta_C > 0$ to C and $\beta_D > 0$ to D. The prize is initially controlled by D. C's prize valuation is common knowledge, while D's is private information

[19] There are three differences between the model here and that of Powell (2015). First, we rule out C's initial option to end the game immediately by accepting the status quo. Any equilibrium in which this occurs involves no choices by D along the path of play, resulting in a trivial direct mechanism for our purposes. The second, related difference is that we normalize the costs k_C and k_D to zero, as both are sunk (and thus decision-irrelevant) once C opts not to accept the status quo. Third, we assume there is no baseline latent risk, i.e., $\underline{r}(p) = 0$ for all p. Our framework could be modified to incorporate non-zero baseline risk at the cost of some additional notation; we omit this possibility here for clarity of exposition.

only known to D. The timing of the game is as follows:

1. Nature selects D's valuation β_D and reveals it to D.

2. C chooses a conventional arms level $p \in [\underline{p}, \overline{p}]$.

3. D chooses a risk level $r \in [0, \overline{r}(p)]$, where $\overline{r}(p) \in (0, 1)$ for each $p \in [\underline{p}, \overline{p}]$.

4. C may quit or continue with the challenge. If C quits, then C receives nothing and D receives the full prize.

5. If C continued with the challenge, D may quit or fight. If D quits, then C receives and full prize and D receives nothing. If instead D fights, then with probability r a nuclear war occurs and each player receives $-n_i < 0$. Otherwise, C wins the prize with probability p and D wins it with probability $1 - p$.

Figure 1 displays the game tree for this model of nuclear brinkmanship.

The baseline framework in section 2 captures some key features of this model, in particular an unintentional risk of nuclear escalation. However, there are also some key differences. The most visible is that D's type now represents their prize valuation instead of their war payoff. Additionally, even after D moves to generate a certain level of endogenous risk, either player may "shut off" that risk by conceding the issue. In this section, we consider a generalized framework that captures the features of the model in Figure 1. We show that the payoff structure in this new framework is essentially isomorphic to a special case in our original contest of capabilities, but that the possibility of shutting off the risk may generate distinct patterns in equilibrium outcomes. Nonetheless, certain outcome patterns can still arise only when risk is generated by a costly political process for D, rather than a costless lever as in the contest of nerves.

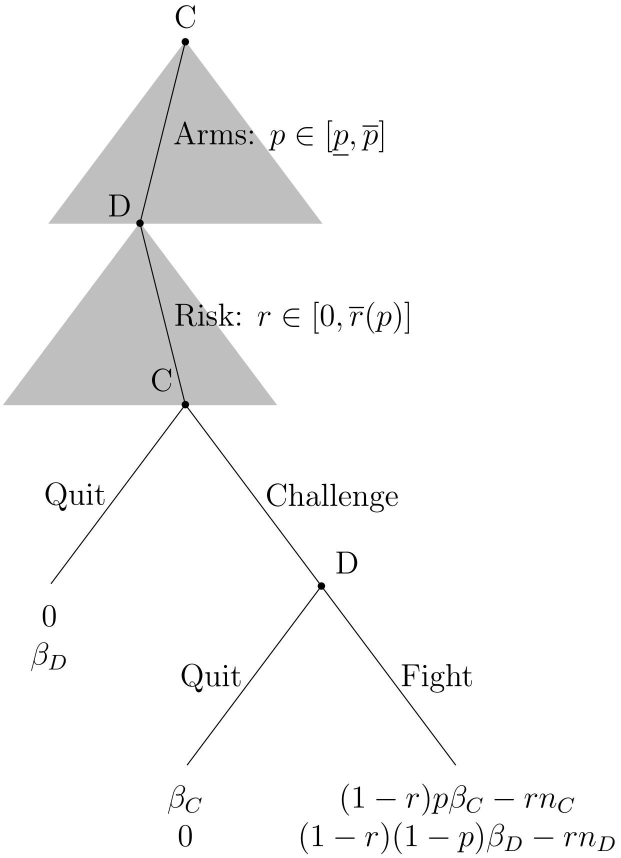
C
Arms: $p \in [\underline{p}, \overline{p}]$
D
Risk: $r \in [0, \overline{r}(p)]$
C
Quit
Challenge
D
0
β_D
Quit
Fight
β_C
0
$(1-r)p\beta_C - rn_C$
$(1-r)(1-p)\beta_D - rn_D$

Figure 1: Game tree for the motivating model where the Defender's type represents their war payoff (based on Powell 2015).

5.1 Direct mechanism

We consider a class of models with the same choice structure as in our baseline framework (introduced in section 2): C and D each choose bargaining actions, including low-level responses, which determine the distribution of spoils and the risk of war. The primitives differ from the baseline model in the following way:

- The war payoffs are fixed values $-n_C, -n_D < 0$.
- The value of the prize for D is a private type, $\beta_D \in [\underline{\beta}_D, \overline{\beta}_D]$.
- The hassling cost function, which we now denote $\kappa_D(h)$, is a function solely of hassling and not of D's type.

The change in the prize payoff structure necessitates a different definition of the direct mechanism than in our baseline analysis. We now model the *share* of the prize received by each type of D, via a function $S_D : [\underline{\beta}_D, \overline{\beta}_D] \rightarrow [0, 1]$. This may represent a share of the prize garnered from negotiations, or from conventional conflict as in the Powell (2015) model. Additionally, mirroring Powell (2015), we allow for the possibility that a state might "quit," ceding the prize to the other while zeroing out the risk of accidental war.[20] We include functions $Q_C, Q_D : [\underline{\beta}_D, \overline{\beta}_D] \rightarrow \{0, 1\}$ to capture these quitting outcomes. For simplicity in the exposition, we restrict $Q_C(\beta_D)+Q_D(\beta_D) \leq 1$; i.e., both states cannot simultaneously quit. To avoid trivialities, we restrict attention to mechanisms in which $Q_C(\beta_D) = Q_D(\beta_D) = 0$ implies $S_D(\beta_D) \in (0, 1)$; i.e., we consider a state to have "quit" if it accepts a settlement with no value. Finally, the assumption of strictly negative war payoffs implies that neither side would start a nuclear war deliberately rather than quit, so we do not incorporate this possibility into the mechanism.

The direct mechanism dictates the payoff to D type β_D of reporting type β'_D. To save space in writing out the reporting function, let $\bar{S}_D(\beta_D)$ denote the final share received by each type

[20] We cannot fully capture this simply by setting $h(\beta_D) = 0$ when a player quits, as D may incur costs of hassling prior to either side quitting.

of D, accounting for either state quitting:

$$\bar{S}_D(\beta_D) \equiv Q_C(\beta_D) + [1 - Q_C(\beta_D) - Q_D(\beta_D)]S_D(\beta_D).$$

This gives us a reporting function of

$$\Psi_D(\beta'_D \mid \beta_D) = \bar{S}_D(\beta'_D)\beta_D - R(h(\beta'_D))n_D - \kappa_D(h(\beta'_D)).$$

[Need to look into the proofs and see whether quitting actually zeroes out the autonomous risk, because this specification of the reporting function makes it look like the autonomous risk stays no matter what]

5.1.1 Application to motivating model

Let us return to the variant of the Powell (2015) brinkmanship model portrayed in Figure 1 above. A pure strategy equilibrium of the game consists of the following quantities and functions. C selects an arms level p^*. Each type of D responds to each possible arms level p with a risk level $r^*(p \mid \beta_D)$. C decides whether to quit after observing arms and risk; let $q^*_C(p, r) \in \{0, 1\}$ be an indicator for C's choice to quit. Finally, if C does not quit, then each type of D chooses whether to do so, denoted $q^*_D(p, r \mid \beta_D) \in \{0, 1\}$.

To adapt this model to our framework, we identify the "hassling" choice as D's selected level of risk. The hassling space is thus $\mathcal{H} = [0, \bar{r}(\bar{p})]$, with associated risk function $R(h) = h$ and cost function $\kappa_D(h) = 0$ (constant). Then given any equilibrium of the game, we can define an equivalent direct mechanism as follows:

- Hassling level: Set $h(\beta_D) = r^*(p^* \mid \beta_D)$ for all β_D.
- C quit: Set $Q_C(\beta_D) = q^*_C(p^*, r^*(p^* \mid \beta_D))$ for all β_D.

- D quit: Set

$$Q_D(\beta_D) = [1 - q_C^*(p^*, r^*(p^* \mid \beta_D))] q_D^*(p^*, r^*(p^* \mid \beta_D) \mid \beta_D)$$

  for all β_D.

- Share of prize if neither quit: Set $S_D(\beta_D) = 1 - p^*$ (constant) for all β_D.

5.1.2 Incentive compatibility and participation constraints

As in the baseline analysis, any direct mechanism corresponding to an equilibrium must satisfy the incentive compatibility condition,

$$\Psi_D(\beta_D \mid \beta_D) \geq \Psi_D(\beta'_D \mid \beta_D) \qquad \text{for all } \beta_D, \beta'_D \in [\underline{\beta}_D, \overline{\beta}_D].$$

In the context of nuclear conflict, a direct mechanism may be incentive compatible yet substantively unlikely. For example, consider a direct mechanism in which $Q_C(\beta_D) = Q_D(\beta_D) = S_D(\beta_D) = 0$ and $h(\beta_D) = \overline{h}$ for all β_D, where $R(\overline{h}) \approx 1$. This trivially satisfies incentive compatibility, but seems unlikely to describe real-world crisis bargaining, as every type of D risks a near-certain nuclear war in order to attain none of the disputed good.

To rule out this sort of implausible equilibrium outcome, we only consider direct mechanisms that satisfy "participation" constraints ensuring that each state prefers the equilibrium over quitting (and, in D's case, not hassling). For D, the participation constraint amounts to each type garnering non-negative expected utility:

$$\Psi_D(\beta_D \mid \beta_D) \geq 0 \qquad \text{for all } \beta_D. \tag{IR-D}$$

It is more complicated to define a participation constraint for C, as the information C has when they decide to quit may vary across game forms. The weakest plausible participation constraint for C is an *ex ante* constraint of non-negative expected utility:

$$\mathbb{E}\left[\bar{S}_C(\beta_D)\beta_C - R(h(\beta_D))n_C\right] \geq 0, \tag{IR-C}$$

where the expectation is taken over the prior distribution of β_D and $\bar{S}_C$ is defined analogously to $\bar{S}_D$. At the other extreme is an *ex post* condition, stating that C yields non-negative expected utility for all type realizations of D:

$$\bar{S}_C(\beta_D)\beta_C - R(h(\beta_D))n_C \geq 0 \qquad \text{for all } \beta_D. \tag{IR-C$'$}$$

5.2 General results

In one sense, reformulating the type space as the prize value rather than as D's war payoff does not change its essential structure or features. Having a higher prize value is akin to having lower war costs or hassling costs, in that it increases D's willingness to run risks in order to achieve a better result at the negotiating table. On its own, then, this reformulation of the type space should not radically change the results of the analysis.

To justify this claim more formally, we can examine the structure of D's payoff function under the modified direct mechanism. A key property of Von Neumann-Morgenstern utility functions is the invariance to affine transformations. If we divide the reporting function $\Psi(\cdot \mid \beta_D)$ by the prize value β_D, we end up with a payoff function that looks akin to the one from our baseline model:

$$\frac{\Psi(\beta_D' \mid \beta_D)}{\beta_D} = \underbrace{\bar{S}_D(\beta_D')}_{V_D(\theta)} - R(h(\beta_D'))\underbrace{\frac{n_D}{\beta_D}}_{\theta} - \underbrace{\frac{\kappa_D(h(\beta_D'))}{\beta_D}}_{K(h,\theta)}.$$

The payoff structure thus should not cause any substantive difference from our baseline contest of capabilities model. This logic is the basis of the next proposition, which states that any equilibrium of the modified model in which neither state quits has an equivalent representation in our original framework.

Proposition 6. *In the model where D's type is prize value, if $Q_C(\beta_D) = Q_D(\beta_D) = 0$ for all β_D, then the direct mechanism is isomorphic to a contest of capabilities in the baseline framework in which Equation 1 is satisfied. If additionally $\kappa_D(h(\beta_D)) = 0$ for all β_D, then it is isomorphic to a contest of nerves.*

An equilibrium of the prize-value model in which neither state ever quits is, in this sense, equivalent to a contest of capabilities that satisfies the decreasing differences condition, Equation 1. Because D would never deliberately provoke war per our assumption that $-n_D < 0$, Proposition 2 then implies that the hassling level and the probability of accidental war weakly increase with D's prize value. Moreover, if risk is generated without any direct cost to D, as in the example model based on Powell (2015), then Proposition 1 and Corollary 1 imply that D's expected utility and settlement value also increase with β_D. From this standpoint, modeling resolve as prize value simply leads to a special (and in fact relatively restrictive) case of the contest of capabilities.

However, Proposition 6 only covers the case in which neither player quits in equilibrium. This rules out certain strategic behaviors related to brinkmanship, where D is prepared to run a high risk of nuclear disaster that is not ultimately realized on the path of play because C would find it intolerable. Once we allow for the possibility of quitting, we obtain a pattern of results that is potentially distinct from the contest of capabilities. Low types of D quit, incurring no costs while receiving nothing. Medium types do not quit but also do not induce C to quit, so any risks they generate are realized on the path of play. The highest types prepare to run a high enough risk—at a strictly greater cost than all low and medium types—to induce C to quit. Essentially, D here can "outbid" C's tolerance for risk, resulting D attaining the prize.

Proposition 7. *In the model where D's type is prize value, there exist $\tilde{\beta}, \hat{\beta} \in [\underline{\beta}_D, \overline{\beta}_D]$ such that:*

(a) For all $\beta_D < \tilde{\beta}$, $Q_D(\beta_D) = 1$ and $\kappa_D(h(\beta_D)) = 0$.

(b) For all $\beta_D \in (\tilde{\beta}, \hat{\beta})$, $Q_C(\beta_D) = Q_D(\beta_D) = 0$. $h(\beta_D)$ and $S_D(\beta_D)$ are weakly increasing on this interval of types.

(c) For all $\beta_D > \hat{\beta}$, $Q_C(\beta_D) = 1$. There exists $\hat{\kappa}$ such that $\kappa_D(h(\beta_D)) < \hat{\kappa}$ for all $\beta_D < \hat{\beta}$ and $\kappa_D(h(\beta_D)) = \hat{\kappa}$ for all $\beta_D > \hat{\beta}$.

In the general setting where type is prize value and the players have the option to quit, we have a nonmonotone ($\cap$-shaped) probability of accidental war as a function of D's type. This is true even though we saw above that the model effectively satisfies the decreasing differences condition (Equation 1), which in our baseline model implied a non-decreasing probability of accidental war among types that do not start a war deliberately.

The nonmonotonicity of the chance of nuclear accidents is a key feature of the model in Schram (2024). But Proposition 7 also shows how it cannot arise in models like our example based on Powell (2015), in which D can generate nuclear risk costlessly. Part (c) of the proposition states that the cost incurred by types that induce C to quit must be *strictly* greater than the costs incurred by types that quit or settle in equilibrium. This cannot happen in a model with costless risk generation—unless all types of D induce C to quit, in which case the probability of accidental war is constant rather than nonmonotone. The upshot is that the nature of the process that generates nuclear risk continues to matter even when we consider a different conceptualization of resolve and allow states to zero out risk by "quitting." Key relationships between resolve and the risk of nuclear accidents arise only when risk is generated by low-level policy options whose attractiveness or capability varies across Defender types.

6 Conclusion

We analyze a class of crisis bargaining models in which states may employ limited policy options short of full-scale war that nonetheless generate a risk of accidentally triggering such a war. Using a mechanism design methodology that allows us to study all equilibria of all such games, we find that their outcomes critically depend on the relationship between a state's private resolve and its willingness and/or ability to use these limited policy instruments. If there is no relationship—i.e., if a state's access to limited conflict is independent of its willingness to engage in full-scale war—then we recover the traditional brinkmanship pattern in which more resolved states engage in more risky limited conflict. However, when the marginal cost of riskier limited policies increases quickly enough with a state's resolve, there are equilibria with the opposite pattern, in which the least resolved types are the most likely to experience accidental escalation. Depending on the technology of limited conflict and its relationship with a state's war payoffs, different bargaining games may lead to completely different patterns of accidental war, even with the same underlying military fundamentals. Our results highlight the complexity of the strategic relationship between resolve, conventional capabilities, and inadvertent escalation.

References

Akçay, Erol, Adam Meirowitz, Kristopher W. Ramsay and Simon A. Levin. 2012. "Evolution of Cooperation and Skew Under Imperfect Information." *Proceedings of the National Academy of Sciences* 109(37):14936–14941.

Ashworth, Scott and Ethan Bueno de Mesquita. 2006. "Monotone comparative statics for models of politics." *American Journal of Political Science* 50(1):214–231.

Banks, Jeffrey S. 1990. "Equilibrium Behavior in Crisis Bargaining Games." *American Journal of Political Science* 34(3):599–614.

Bas, Muhammet A and Andrew J Coe. 2012. "Arms diffusion and war." *Journal of Conflict Resolution* 56(4):651–674.

Brodie, Bernard. 1966. *Escalation and the Nuclear Option.* Princeton University Press.

Cobzaş, Ştefan, Radu Miculescu and Adriana Nicolae. 2019. *Lipschitz Functions.* Springer.

Fey, Mark and Brenton Kenkel. 2021. "Is an Ultimatum the Last Word on Crisis Bargaining?" *Journal of Politics* 83(1):87–102.

Fey, Mark and Kristopher W. Ramsay. 2009. "Mechanism Design Goes to War: Peaceful Outcomes with Interdependent and Correlated Types." *Review of Economic Design* 13(3):233.

Fey, Mark and Kristopher W Ramsay. 2011. "Uncertainty and Incentives in Crisis Bargaining: Game-Free Analysis of International Conflict." *American Journal of Political Science* 55(1):149–169.

Jervis, Robert. 1976. *Perception and misperception in international politics*. Princeton University Press.

Jervis, Robert. 1979. "Why nuclear superiority doesn't matter." *Political Science Quarterly* 94(4):617–633.

Kenkel, Brenton and Peter Schram. 2024. "Uncertainty in Crisis Bargaining with Multiple Policy Options." Forthcoming in *American Journal of Political Science*.
URL: https://doi.org/10.1111/ajps.12849

Liu, Linqun. 2021. "Domestic Constraints in Crisis Bargaining." Typescript, University of Chicago.
URL: https://sites.google.com/site/liqunliu90/home

Liu, Liqun et al. 2021. Domestic Constraints in Crisis Bargaining. Technical report.

Milgrom, Paul and Chris Shannon. 1994. "Monotone comparative statics." *Econometrica: Journal of the Econometric Society* pp. 157–180.

Milgrom, Paul and Ilya Segal. 2002. "Envelope Theorems for Arbitrary Choice Sets." *Econometrica* 70(2):583–601.

Myerson, Roger B. 1979. "Incentive Compatibility and the Bargaining Problem." *Econometrica* 47(1):61–73.

Nalebuff, Barry. 1986. "Brinkmanship and nuclear deterrence: The neutrality of escalation." *Conflict Management and Peace Science* 9(2):19–30.

Paul, Derek, Michael D Intriligator, Paul Smoker et al. 1990. *Accidental Nuclear War: Proceedings of the Eighteenth Pugwash Workshop on Nuclear Forces*. Dundurn.

Perrow, Charles. 2011. *Normal accidents*. Princeton university press.

Posen, Barry R. 2014. *Inadvertent escalation*. Cornell University Press.

Powell, Robert. 1988. "Nuclear brinkmanship with two-sided incomplete information." *American Political Science Review* 82(1):155–178.

Powell, Robert. 1989. "Nuclear Deterrence and the Strategy of Limited Retaliation." *American Political Science Review* 83(2):503–519.

Powell, Robert. 1990. *Nuclear deterrence theory: The search for credibility.* Cambridge University Press.

Powell, Robert. 2015. "Nuclear Brinkmanship, Limited War, and Military Power." *International Organization* 69(3):589–626.

Sagan, Scott D. 1994. "The perils of proliferation: Organization theory, deterrence theory, and the spread of nuclear weapons." *International Security* 18(4):66–107.

Schelling, Thomas C. 1966. *Arms and influence.* Yale University Press.

Schelling, Thomas C. 1980. *The strategy of conflict.* Harvard university press.

Schram, Peter. 2024. "Conflicts that Leave Something to Chance: Establishing Brinkmanship through Conventional Wars." *Working Paper* .

Snyder, Glenn H. 1965. The Balance of Power and the Balance of Terror. In *World in Crisis*, ed. Fredrick H. Hartmann. The Macmillan Company chapter 20, pp. 180–191.

Spaniel, William. 2019. *Bargaining over the Bomb: The Successes and Failures of Nuclear Negotiations.* Cambridge University Press.

Spaniel, William. 2020. "Power Transfers, Military Uncertainty, and War." *Journal of Theoretical Politics* 32(4):538–556.

Appendix

Contents

A Simple Brinkmanship Game

A.1 Assumptions

Note that in this game, for a selected risk level $r \in \{\underline{r}, \bar{r}\}$ a type $v_D \in \{\underline{v}_D, \bar{v}_D\}$ D will escalate whenever

$$(1-r)\left(pv_D - k_D\right) + r\left(-n_D - k_D\right) \geq -\, k_D,$$

or more simply

$$(1-r)pv_D - rn_D \geq 0,$$

and will quit otherwise. We assume that if C challenges after observing $\bar{r}$, then type $\underline{v}_D$ Ds will quit and type $\bar{v}_D$Ds will fight. Using the above, formally, this is $(1-\bar{r})p\underline{v}_D - \bar{r}n_D < 0$ and $(1-\bar{r})p\bar{v}_D - \bar{r}n_D > 0$ (respectively). We also assume that if C challenges after observing $\underline{r}$, that both types of D prefer fighting to acquiescing, which is $(1-\underline{r})p\underline{v}_D - \underline{r}n_D > 0$.

We also assume that if D sets $r = \underline{r}$, then C prefers to challenge, even if doing so always results in a fight. Formally, this is

$$(1-\underline{r})\left((1-p)v_C - k_C\right) + \underline{r}\left(-n_C - k_C\right) > -k_C$$

or more simply

$$(1-\underline{r})(1-p)v_C + \underline{r}n_C > 0.$$

A.2 Equilibrium Statement

The following constitutes a perfect Bayesian Nash Equilibrium. Note that by the assumptions above, Stage 4 equilibrium behavior is the same across all cases: if $r = \underline{r}$ and C challenges, then both types of D will fight; and, if $r = \bar{r}$ and C challenges, then type $\bar{v}_D$ will fight and type $\underline{v}_D$ will acquiece.

A.2.1 Case A, $(1-\zeta)\left((1-\bar{r})(1-p)v_C - \bar{r}n_C\right) + \zeta v_C \leq 0$,

- Stage 3: If $r = \underline{r}$, then C will challenge and believes D is high-resolve with probability 1. If $r = \bar{r}$, then C does not challenge and believes D is low-resolve with probability ζ and high-resolve with probability $1-\zeta$.
- Stage 2: Both types of D will set $r = \bar{r}$.

A.2.2 Case B, $(1-\bar{r})(1-p)v_C - \bar{r}n_C > 0$

- Stage 3: If $r = \underline{r}$, then C will challenge and believes D is low-resolve with probability ζ and high-resolve with probability $1-\zeta$. If $r = \bar{r}$, then C will challenge and believes D is high-resolve with probability 1.

- Stage 2: Both types of D will set $r = \underline{r}$.

A.2.3 Case C, $(1-\zeta)\left((1-\bar{r})(1-p)v_C - \bar{r}n_C\right) + \zeta v_C > 0$, $(1-\bar{r})(1-p)v_C - \bar{r}n_C \leq 0$, and $(1 - r$

)**p**$\bar{v}_D - r$

$\mathbf{n}_D > (1-\beta)\bar{v}_D + \beta\left((1-\bar{r})p\bar{v}_D - \bar{r}n_D\right)$

- Stage 3: If $r = \underline{r}$, then C will challenge and believes D is low-resolve with probability ζ and high-resolve with probability $1-\zeta$. If $r = \bar{r}$, then believes D is low-resolve with probability 1.
- Stage 2: Both types of D will set $r = \underline{r}$.

A.2.4 Case D, otherwise...

- Stage 3: If $r = \underline{r}$, then C will challenge and believes D is low-resolve with probability 1. If $r = \bar{r}$, then C will challenge with probability $\beta = \frac{\underline{v}_D - (1-\underline{r})p\underline{v}_D + \underline{r}n_D}{\underline{v}_D}$ and not challenge with probability $1-\beta$; also, observing $r = \bar{r}$, C believes D is high-resolve with probability $\mu = \frac{v_C}{v_C + \bar{r}n_C - (1-\bar{r})(1-p)v_C}$ and low-resolve with probability $1-\mu$.
- Stage 2: High-resolve D always sets $r = \bar{r}$. Low-resolve D sets $r = \bar{r}$ with probability $\alpha = \frac{(1-\zeta)(\bar{r}n_C - (1-\bar{r})(1-p)v_C)}{\zeta v_C}$ and sets $r = \underline{r}$ with probability $1-\alpha$.

A.3 Proof of Equilibrium

In all cases, Stage 4 behavior is optimal following the assumptions.

Case A: The condition $((1-\zeta)\left((1-\bar{r})(1-p)v_C - \bar{r}n_C\right) + \zeta v_C \leq 0)$ implies that, under high nuclear risk$\bar{r}$, C weakly prefers acquiescing to challenging when high-resolve types will fight and low-resolve types will quit when challenged. By assumption, C would challenge if $r = \underline{r}$. C's on-path beliefs follow priors. Both types of D prefer setting risk $\bar{r}$ and attaining the asset without a fight to setting risk $\underline{r}$and fighting over the asset. The following parameter values support the condition of the case and the assumptions: $\bar{v}_D = 1$, $\underline{v}_D = 0.2$, $n_D = 1.5$, $k_D = 0.1$, $v_C = 1$, $n_C = 3$, $k_C = 0.1$, $\underline{r} = 0.05$, $\bar{r} = 0.2$, $p = 0.8$, and $\zeta = 0.2$.

Case B: The condition $(1-\bar{r})(1-p)v_C - \bar{r}n_C > 0$ implies that, under high nuclear risk$\bar{r}$, C strictly prefers challenging to acquiescing when challenging always results in a fight. By assumption, C will challenge if $r = \underline{r}$. C's on-path beliefs follow priors. Given that C will always challenge, both types of D do best selecting the low-nuclear risk level and fighting over the asset to setting the high-nuclear risk level and fighting (for high-resolve types) or acquiescing (for low-resolve types) after being challenged. The following parameter values support the condition of the case and the assumptions: $\bar{v}_D = 1$, $\underline{v}_D = 0.2$, $n_D = 1.5$, $k_D = 0.1$, $v_C = 2$, $n_C = 1$, $k_C = 0.1$, $\underline{r} = 0.05$, $\bar{r} = 0.2$, $p = 0.8$, and $\zeta = 0.5$.

Case C: The condition $(1-\zeta)\left((1-\bar{r})(1-p)v_C - \bar{r}n_C\right) + \zeta v_C > 0$ implies that, under high nuclear risks $\bar{r}$, C prefers challenging to acquiescing when low- and high-resolve types always select $\bar{r}$ and follow Stage 4 equilibrium behavior. The condition $(1-\bar{r})(1-p)v_C - \bar{r}n_C \leq 0$ implies that, under high nuclear risk $\bar{r}$, C prefers acquiescing to challenging when only high-resolve types set $\bar{r}$ and always fight. The last condition $((1-\underline{r})p\bar{v}_D - \underline{r}n_D > (1-\beta)\bar{v}_D + \beta\left((1-\bar{r})p\bar{v}_D - \bar{r}n_D\right))$ implies that a semi-separating equilibrium cannot exist (and is elaborated on in the discussions on Case D). By assumption, C will challenge if $r = \underline{r}$. C would also prefer to challenge upon observing $(r = \bar{r})$ given C believes that only low-resolve D would ever select $r = \bar{r}$ and would then quit when challenged. C's on-path beliefs follow priors. Given C's behavior, both types of D prefer setting $\underline{r}$ and fighting over the asset to setting $\bar{r}$ and fighting over the asset. The following parameter values support the condition of the case and the assumptions: $\bar{v}_D = 1$, $\underline{v}_D = 0.2$, $n_D = 1.5$, $k_D = 0.1$, $v_C = 1$, $n_C = 2$, $k_C = 0.1$, $\underline{r} = 0.05$, $\bar{r} = 0.2$, $p = 0.8$, and $\zeta = 0.5$.

Case D: The conditions $(1-\zeta)\left((1-\bar{r})(1-p)v_C - \bar{r}n_C\right) + \zeta v_C > 0$ and $(1-\bar{r})(1-p)v_C - \bar{r}n_C \leq 0$ were discussed in Case C. Let μ denote the probability that D is high-resolve after C observes D play $r = \bar{r}$ (formally, $Pr(v_D = \bar{v}_D | r = \bar{r}) = \mu$). Observing $\bar{r}$, C will be indifferent between challenging (anticipating that high-resolve D will fight and low-resolve D will quit) and not when

$$\mu\left((1-\bar{r})(1-p)v_C - \bar{r}n_C\right) + (1-\mu)v_C = 0,$$

or

$$\mu = \frac{v_C}{v_C + \bar{r}n_C - (1-\bar{r})(1-p)v_C}.$$

Using Bayes' rule, we can alternatively define μ in terms of both types of D's actions. We define

$$\mu = Pr(v_D = \bar{v}_D | r = \bar{r}) = \frac{Pr(r = \bar{r} | v_D = \bar{v}_D) * Pr(v_D = \bar{v}_D)}{Pr(r = \bar{r} | v_D = \bar{v}_D) * Pr(v_D = \bar{v}_D) + Pr(r = \bar{r} | v_D = \underline{v}_D) * Pr(v_D = \underline{v}_D)},$$

which is equivalent to

$$\mu = \frac{(1-\zeta)}{(1-\zeta) + \alpha\zeta},$$

where α denotes the probability that type $\underline{v}_D$ sets $r = \bar{r}$. Combining the two expressions for μ allows us to characterize α as

$$\alpha = \frac{(1-\zeta)\left(\bar{r}n_C - (1-\bar{r})(1-p)v_C\right)}{\zeta v_C},$$

suggesting that so long that type $\underline{v}_D$ sets $r = \bar{r}$ with probability α, C will be indifferent between challenging and not challening upon observing $\bar{r}$. This indifference supports C's mixing behavior upon observing $\bar{r}$.

Additionally, by assumption, C would challenge if $r = \underline{r}$, and C's on-path beliefs are calculated via Bayes rule.

A low-resolve D will be indifferent between setting $r = \underline{r}$ and always fighting and setting $r = \bar{r}$ with C sometimes challenging and sometimes not challenging when, letting β denote the likelihood that C challenges after observing $r = \bar{r}$, the following holds:

$$(1-\underline{r})p\underline{v}_D - \underline{r}n_D = (1-\beta)\underline{v}_D.$$

This can be re-written as

$$\beta = \frac{\underline{v}_D - (1-\underline{r})p\underline{v}_D + \underline{r}n_D}{\underline{v}_D}.$$

This indifference supports low-resolve D's mixing behavior between setting $r = \underline{r}$ and $r = \bar{r}$.

Finally, we must show that high-resolve D is willing to always select $r = \bar{r}$. High-resolve D prefers $r = \bar{r}$ and sometimes fighting to setting $r = \underline{r}$ and always fighting whenever $(1-\underline{r})p\bar{v}_D - \underline{r}n_D \leq (1-\beta)\bar{v}_D + \beta\left((1-\bar{r})p\bar{v}_D - \bar{r}n_D\right)$; this is implied by the conditions of the case.

The following parameter values support the condition of the case and the assumptions: $\bar{v}_D = 1$, $\underline{v}_D = 0.2$, $n_D = 1.5$, $k_D = 0.1$, $v_C = 1$, $n_C = 2$, $k_C = 0.1$, $\underline{r} = 0.05$, $\bar{r} = 0.1$, $p = 0.8$, and $\zeta = 0.5$.

A.4 Proof that High-Resolve D Risks Nuclear War (Weakly) More

We claim that high-resolve Ds risk a nuclear exchange with greater probability than low-resolve Ds.

This holds trivially in Cases A, B, and C. For Case D, we can say the following for low-resolve types:

$$(1-\underline{r})p\underline{v}_D - \underline{r}n_D = (1-\beta)\underline{v}_D$$

$$(1-\underline{r})p\underline{v}_D - \underline{r}n_D > (1-\beta)\underline{v}_D + \beta\left((1-\bar{r})p\underline{v}_D - \bar{r}n_D\right)$$

$$n_D\left(\beta\bar{r} - \underline{r}\right) > \underline{v}_D\left((1-\beta) + \beta(1-\bar{r})p - (1-\underline{r})p\right) \tag{A.1}$$

The first line holds based on how β is defined. The second line holds because low-resolve D prefers acquiescing to challlenging under $\bar{r}$. The last line is algebra.

Also, for high-resolve types, we can say the following:

$$\begin{aligned}(1-\underline{r})p\bar{v}_D - \underline{r}n_D \leq & (1-\beta)\bar{v}_D + \beta\left((1-\bar{r})p\bar{v}_D - \bar{r}n_D\right) \\ n_D\left(\beta\bar{r} - \underline{r}\right) \leq & \bar{v}_D\left((1-\beta) + \beta(1-\bar{r})p - (1-\underline{r})p\right) \qquad \text{(A.2)}\end{aligned}$$

The first line holds based on the conditions of the case, and the second line is algebra.

Because $\underline{v}_D < \bar{v}_D$, combining A.1 and A.2 implies that $(1-\beta)+\beta(1-\bar{r})p-(1-\underline{r})p$ must be positive, which also implies that $\underline{r} < \beta\bar{r}$. In Case D, the likelihood that type $\underline{v}_D$ enters into a nuclear exchange is $(1-\alpha)\underline{r}$, and the likelihood that type $\bar{v}_D$ enters into a nuclear exchange is $\beta\bar{r}$. Because $\alpha \in [0,1]$ and $\underline{r} < \beta\bar{r}$, high-resolve D's risks nuclear war with higher likelihood.

B Simple Contest of Capabilities Game

B.1 Assumptions

Similar to the brinkmanship game, we assume that $(1-\bar{r})p\underline{v}_D - \bar{r}n_D < 0$, $(1-\bar{r})p\bar{v}_D - \bar{r}n_D > 0$, $(1-\underline{r})p\underline{v}_D - \underline{r}n_D > 0$, and $(1-\underline{r})(1-p)v_C + \underline{r}n_C > 0$. We now also assume that that high-resolve Ds prefer selecting the high nuclear escalation level (with high sunk costs) if this means that they attain the asset without a fight relative to selecting the low nuclear escalation level (with no sunk costs) and then fighting over the asset. Formally, this is the following:

$$\bar{v}_D - K(\bar{r}) > (1-\underline{r})p\bar{v}_D - \underline{r}n_D$$

We also assume that low-resolve types have the opposite preferences: they prefer selecting into the low nuclear escalation level and fighting over the asset to selecting the high nuclear escalation level and incurring the high sunk costs, even if it means that they attain the asset. Formally, this is the following:

$$\underline{v}_D - K(\bar{r}) < (1-\underline{r})p\underline{v}_D - \underline{r}n_D$$

B.2 Equilibrium Statement

The following constitutes a perfect Bayesian Nash Equilibrium. Note that by the assumptions above, Stage 4 equilibrium behavior is the same across all cases: if $r = \underline{r}$ and C challenges, then both types of D will fight; and, if $r = \bar{r}$ and C challenges, then type $\bar{v}_D$ will fight and type $\underline{v}_D$ will acquiece.

B.2.1 Case E, $0 < (1-\bar{r})(1-p)v_C - \bar{r}n_C$,

- Stage 3: If $r = \underline{r}$, then C will challenge and believes D is low-resolve with probability ζ and high-resolve with probability $1-\zeta$. If $r = \bar{r}$, then C will challenge and believes

D is high-resolve with probability 1.

- Stage 2: Both types of D will set $r = \underline{r}$.

B.2.2 Case F, otherwise...

- Stage 3: If $r = \underline{r}$, then C will challenge and believes D is low-resolve with probability 1. If $r = \bar{r}$, then C will not challenge and believes D is high-resolve with probability 1.
- Stage 2: High-resolve Ds will set $r = \bar{r}$ and low-resolve Ds will set $r = \underline{r}$.

B.3 Proof of Equilibrium

In all cases, Stage 4 behavior is optimal following the assumptions.

Case E: The condition $0 < (1-\bar{r})(1-p)v_C - \bar{r}n_C$ implies that, under high nuclear risk $\bar{r}$, C prefers challenging to acquiescing when C must fight after challenging. By assumption, C would challenge if $r = \bar{r}$ or $r = \underline{r}$. C's on-path beliefs follow priors. Both types of D prefer setting risk $\underline{r}$ and fighting over the asset to setting risk $\bar{r}$ and fighting over the asset (for high-resolve types) or acquiescing (for low-resolve types). The following parameter values support the condition of the case and the assumptions: $\bar{v}_D = 1.5$, $\underline{v}_D = 0.2$, $n_D = 1.5$, $k_D = 0.1$, $v_C = 1.5$, $n_C = 2$, $k_C = 0.1$, $\underline{r} = 0.05$, $\bar{r} = 0.1$, $p = 0.8$, $\zeta = 0.5$, and $K(\bar{r}) = 0.4$.

Case F: The condition $0 \geq (1-\bar{r})(1-p)v_C - \bar{r}n_C$ implies that, under high nuclear risk $\bar{r}$, C prefers acquiescing to challenging when C must fight after challenging. By assumption, C would challenge if $r = \underline{r}$ and would acquiesce if $r = \bar{r}$. C's on-path beliefs follow D's strategic play. The assumption $\bar{v}_D - K(\bar{r}) > (1-\underline{r})p\bar{v}_D - \underline{r}n_D$ implies that high-resolve D prefers setting $r = \bar{r}$ and having C acquiesce to setting $r = \underline{r}$ and fighting over the asset. In contrast, the assumption $\underline{v}_D - K(\bar{r}) < (1-\underline{r})p\underline{v}_D - \underline{r}n_D$ implies that low-resolve D prefers setting $r = \underline{r}$ and fighting over the asset to setting $r = \bar{r}$ and having C acquiesce. The following parameter values support the condition of the case and the assumptions: $\bar{v}_D = 1.5$, $\underline{v}_D = 0.2$, $n_D = 1.5$, $k_D = 0.1$, $v_C = 1$, $n_C = 2$, $k_C = 0.1$, $\underline{r} = 0.05$, $\bar{r} = 0.1$, $p = 0.8$, $\zeta = 0.5$, and $K(\bar{r}) = 0.4$.

C Proofs of Named Results

C.1 Proof of Proposition 1

Proposition 1. *In any equilibrium of a contest of nerves, the total probability of war and the Defender's equilibrium utility weakly increase with the Defender's resolve: if $\theta' < \theta''$, then $\pi(\theta')[1 - R(h(\theta'))] \geq \pi(\theta'')[1 - R(h(\theta''))]$ and $U(\theta') \leq U(\theta'')$.*

Proof. We prove the result by showing that there is a payoff-equivalent direct mechanism in an ordinary crisis bargaining game that satisfies the incentive compatibility conditions of

Banks (1990). For Banks (1990), a direct mechanism is defined by a function $x : \Theta \to \mathbb{R}$ giving settlement values and a function $p : \Theta \to [0,1]$ giving the probability of war. Given such a mechanism, the expected utility to type θ for mimicking the bargaining strategy of type θ' is given by

$$\tilde{\Phi}_D(\theta' \mid \theta) = p(\theta')\theta + (1 - p(\theta'))x(\theta').$$

Now consider a direct mechanism for a contest of nerves that satisfies our incentive compatibility condition, (IC), and define the following functions:

$$\begin{aligned} x(\theta) &= V_D(\theta) - \frac{\kappa_D(h(\theta))}{1 - R(h(\theta))}, \\ p(\theta) &= 1 - \pi(\theta)\left[1 - R(h(\theta))\right]. \end{aligned}$$

For all $\theta, \theta' \in \Theta$, we have

$$\begin{aligned} \tilde{\Phi}_D(\theta' \mid \theta) &= p(\theta')\theta + (1 - p(\theta'))x(\theta') \\ &= \left(1 - \left[1 - R(h(\theta'))\right]\pi(\theta')\right)\theta + \pi(\theta')\left[1 - R(h(\theta'))\right]V_D(\theta') - \pi(\theta')\kappa_D(h(\theta')) \\ &= \pi(\theta')\left[(1 - R(h(\theta')))V_D(\theta') + R(h(\theta'))\theta - \kappa_D(h(\theta'))\right] + (1 - \pi(\theta'))\theta \\ &= \Phi_D(\theta' \mid \theta). \end{aligned}$$

Incentive compatibility of the original direct mechanism therefore implies incentive compatibility of the Banks (1990) mechanism (x, p). The first claim of the proposition then follows from Lemma 1 of Banks (1990), and the second follows from his Lemma 4. □

C.2 Proof of Corollary 1

Corollary 1. *In any equilibrium of a contest of nerves, if war never occurs deliberately ($\pi(\theta) = 1$ for all θ), then the probability of accidental war and the Defender's settlement value weakly increase with the Defender's resolve: if $\theta' < \theta''$, then $R(h(\theta')) \leq R(h(\theta''))$ and $V_D(\theta') \leq V_D(\theta'')$.*

Proof. The first claim is immediate from Proposition 1, setting $\pi(\theta') = \pi(\theta'') = 1$. To prove the second claim, observe that the function $\frac{\kappa_D}{1-R}$ is weakly increasing in h, as κ_D and R are both non-decreasing in h. Following the proof of Proposition 1, Lemma 2 of Banks (1990) implies

$$V_D(\theta'') - V_D(\theta') \geq \frac{\kappa_D(h(\theta''))}{1 - R(h(\theta''))} - \frac{\kappa_D(h(\theta'))}{1 - R(h(\theta'))}.$$

Because R is strictly increasing, $R(h(\theta'')) \geq R(h(\theta'))$ implies $h(\theta'') \geq h(\theta')$, so the RHS of the above expression is non-negative. □

C.3 Proof of Proposition 2

Proposition 2. *Consider an equilibrium of a contest of capabilities.*

(a) *The probability of accidental war weakly increases with the Defender's resolve if*

$$[K_D(h'',\theta'') - K_D(h',\theta'')] - [K_D(h'',\theta') - K_D(h',\theta')] < [R(h'') - R(h')](\theta'' - \theta') \quad (1)$$

for all $h', h'' \in h(\Theta_1)$ *and* $\theta', \theta'' \in \Theta_1$ *such that* $h' < h''$ *and* $\theta' < \theta''$.

(b) *The probability of accidental war weakly decreases with the Defender's resolve if*

$$[K_D(h'',\theta'') - K_D(h',\theta'')] - [K_D(h'',\theta') - K_D(h',\theta')] > [R(h'') - R(h')](\theta'' - \theta') \quad (2)$$

for all $h', h'' \in h(\Theta_1)$ *and* $\theta', \theta'' \in \Theta_1$ *such that* $h' < h''$ *and* $\theta' < \theta''$.

Proof. We will prove the first claim; the proof of the second is analogous. Take any $\theta', \theta'' \in \Theta_1$ such that $\theta' < \theta''$, and suppose $h(\theta') > h(\theta'')$. To economize on notation in the remainder of the proof, define $h' \equiv h(\theta')$, $V' \equiv V_D(\theta')$, and $R' \equiv R(h(\theta'))$; and let h'', V'', and R'' be defined analogously. Incentive compatibility for θ' implies

$$(1 - R')V' + R'\theta' - K_D(h',\theta') \geq (1 - R'')V'' + R''\theta' - K_D(h'',\theta'),$$

which is equivalent to

$$(R' - R'')\theta' - [K_D(h',\theta') - K_D(h'',\theta')] \geq (1 - R'')V'' - (1 - R')V'.$$

Similarly, incentive compatibility for θ'' implies

$$(1 - R'')V'' + R''\theta'' - K_D(h'',\theta'') \geq (1 - R')V' + R'\theta'' - K_D(h',\theta''),$$

which is equivalent to

$$(1 - R'')V'' - (1 - R')V' \geq (R' - R'')\theta'' - [K_D(h',\theta'') - K_D(h'',\theta'')].$$

Combining the incentive compatibility conditions and rearranging terms gives

$$[K_D(h',\theta'') - K_D(h'',\theta'')] - [K_D(h',\theta') - K_D(h'',\theta')] \geq (R' - R'')(\theta'' - \theta').$$

Because $h' > h''$ and $\theta'' > \theta'$, this implies that Equation 1 does not hold. □

C.4 Proof of Proposition 3

The proof of the proposition follows a series of lemmas. The method of proof is similar to other envelope theorem analyses (e.g., in Banks 1990; Kenkel and Schram 2024).

C.4.1 Envelope theorem

Lemma C.1. *Suppose the differentiability assumptions hold. For any IC direct mechanism in which all types settle, we have*

$$U_D(\theta) = U(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \left[R(h(t)) - \frac{\partial K_D(h(t), t)}{\partial t} \right] dt \tag{C.3}$$

for all $\theta \in \Theta$.

Proof. (IC) implies $U_D(\theta) = \sup_{\theta' \in \Theta} \Phi_D(\theta' \mid \theta)$ for all $\theta \in \Theta$. The differentiability assumptions imply that $\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta}$ exists for all $\theta, \theta' \in \Theta$. Corollary 1 of Milgrom and Segal (2002) then implies

$$\begin{aligned} U_D(\theta) &= U_D(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \left. \frac{\partial \Phi_D(\theta' \mid t)}{\partial t} \right|_{\theta' = t} dt \\ &= U_D(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \left[R(h(t)) - \frac{\partial K_D(h(t), t)}{\partial t} \right] dt \end{aligned}$$

for all $\theta \in \Theta$, as claimed. □

C.4.2 Value of settlement

Lemma C.2. *Suppose the differentiability assumptions hold, and consider a direct mechanism in which $\pi = 1$. For all $\theta \in \Theta$, Equation C.3 is satisfied if and only if*

$$V_D(\theta) = \frac{U(\underline{\theta}) - R(h(\theta))\theta + K_D(h(\theta), \theta) + \int_{\underline{\theta}}^{\theta} \left[R(h(t)) - \frac{\partial K_D(h(t),t)}{\partial t} \right] dt}{1 - R(h(\theta))}. \tag{C.4}$$

Proof. Immediate from setting

$$(1 - R(h(\theta)))V_D(\theta) + R(h(\theta))\theta - K_D(h(\theta), \theta) = U_D(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \left[R(h(t)) - \frac{\partial K_D(h(t), t)}{\partial t} \right] dt$$

and solving for $V_D(\theta)$. □

C.4.3 Global incentive compatibility

Lemma C.3. *Suppose the differentiability assumptions hold. Consider a direct mechanism in which $\pi = 1$, $V_D(\theta)$ satisfies Equation C.4 for all $\theta \in \Theta$, and h is absolutely continuous. For all $\theta \in \Theta$, $\Phi_D(\cdot \mid \theta)$ is differentiable almost everywhere, with*

$$\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} = h'(\theta') \left[R'(h(\theta'))(\theta - \theta') - \left(\frac{\partial K_D(h(\theta'), \theta)}{\partial h} - \frac{\partial K_D(h(\theta'), \theta')}{\partial h} \right) \right] \tag{C.5}$$

at each point of differentiability.

Proof. For all $\theta, \theta' \in \Theta$,

$$\begin{aligned}\Phi_D(\theta' \mid \theta) &= (1 - R(h(\theta')))V_D(\theta') + R(h(\theta'))\theta - K_D(h(\theta'), \theta) \\ &= (1 - R(h(\theta')))V_D(\theta') + R(h(\theta'))\theta' - K_D(h(\theta'), \theta') \\ &\quad + R(h(\theta'))(\theta - \theta') - [K_D(h(\theta'), \theta) - K_D(h(\theta'), \theta' \\ &= U_D(\theta') + R(h(\theta'))(\theta - \theta') - [K_D(h(\theta'), \theta) - K_D(h(\theta'), \theta')].\end{aligned}$$

As R and K_D are Lipschitz (via their continuous differentiability), absolute continuity of h implies $R(h(\cdot))$ and $K_D(h(\cdot), \cdot)$ are absolutely continuous and thus differentiable almost everywhere (Cobzaş, Miculescu and Nicolae 2019). Additionally, Lemma C.2 implies that U_D is absolutely continuous as well. Therefore, $\Phi_D(\cdot \mid \theta)$ is differentiable almost everywhere, with

$$\begin{aligned}\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} &= U_D'(\theta') + R'(h(\theta'))h'(\theta')(\theta - \theta') - R(h(\theta')) - \frac{\partial K_D(h(\theta'), \theta)}{\partial h} h'(\theta') \\ &\quad + \frac{\partial K_D(h(\theta'), \theta')}{\partial h} h'(\theta') + \frac{\partial K_D(h(\theta'), \theta')}{\partial \theta'} \\ &= h'(\theta') \left[R'(h(\theta'))(\theta - \theta') - \left(\frac{\partial K_D(h(\theta'), \theta)}{\partial h} - \frac{\partial K_D(h(\theta'), \theta')}{\partial h} \right) \right]\end{aligned}$$

at each point of differentiability. □

C.4.4 Proof of proposition

Proposition 3. *Suppose the differentiability assumptions hold.*

(a) Let h^* be any absolutely continuous, weakly increasing function that satisfies

$$\frac{\partial^2 K_D(h, \theta)}{\partial h \partial \theta} \leq R'(h) \qquad \text{for all } h \in h^*(\Theta),\ \theta \in \Theta.$$

There is an incentive compatible direct mechanism in which $\pi(\theta) = 1$ and $h(\theta) = h^*(\theta)$ for all $\theta \in \Theta$.

(b) Let h^{**} be any absolutely continuous, weakly decreasing function that satisfies

$$\frac{\partial^2 K_D(h, \theta)}{\partial h \partial \theta} \geq R'(h) \qquad \text{for all } h \in h^{**}(\Theta), \theta \in \Theta.$$

There is an incentive compatible direct mechanism in which $\pi(\theta) = 1$ and $h(\theta) = h^{**}(\theta)$ for all $\theta \in \Theta$.

Proof. We prove the first claim; the proof of the second is analogous. Take any $V_0 \in \mathbb{R}$, set $V_D(\underline{\theta}) = V_0$,[1] and then define $V_D(\theta)$ according to Equation C.4 for all $\theta \in (\underline{\theta}, \overline{\theta}]$. For any

[1] A task for future work is to identify a sharp condition under which (VA) holds as well.

$\theta \in \Theta$ and almost all $\theta' \in [\underline{\theta}, \theta)$, Lemma C.3 implies

$$\begin{aligned}
\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} &= h'(\theta') \left[R'(h(\theta'))(\theta - \theta') - \left(\frac{\partial K_D(h(\theta'), \theta)}{\partial h} - \frac{\partial K_D(h(\theta'), \theta')}{\partial h} \right) \right] \\
&= h'(\theta') \left[R'(h(\theta'))(\theta - \theta') - \int_{\theta'}^{\theta} \frac{\partial^2 K_D(h(\theta'), t)}{\partial h \partial t} \, dt \right] \\
&\geq h'(\theta') \left[R'(h(\theta'))(\theta - \theta') - \int_{\theta'}^{\theta} R'(h(\theta')) \, dt \right] \\
&= 0.
\end{aligned}$$

Therefore, $\Phi_D(\theta \mid \theta) \geq \Phi_D(\theta' \mid \theta)$ for all $\theta' < \theta$. Similarly, for almost all $\theta' \in (\theta, \overline{\theta}]$,

$$\begin{aligned}
\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} &= h'(\theta') \left[R'(h(\theta'))(\theta - \theta') - \left(\frac{\partial K_D(h(\theta'), \theta)}{\partial h} - \frac{\partial K_D(h(\theta'), \theta')}{\partial h} \right) \right] \\
&= h'(\theta') \left[\left(\frac{\partial K_D(h(\theta'), \theta')}{\partial h} - \frac{\partial K_D(h(\theta'), \theta)}{\partial h} \right) - R'(h(\theta'))(\theta' - \theta) \right] \\
&= h'(\theta') \left[\int_{\theta}^{\theta'} \frac{\partial^2 K_D(h(\theta'), t)}{\partial h \partial t} \, dt - R'(h(\theta'))(\theta' - \theta) \right] \\
&\leq h'(\theta') \left[\int_{\theta}^{\theta'} R'(h(\theta')) \, dt - R'(h(\theta'))(\theta' - \theta) \right] \\
&= 0.
\end{aligned}$$

Therefore, $\Phi_D(\theta \mid \theta) \geq \Phi_D(\theta' \mid \theta)$ for all $\theta' > \theta$, which combined with the prior result implies that the direct mechanism satisfies (IC). □

C.5 Proof of Proposition 4

Proposition 4. *Consider an equilibrium of a contest of capabilities. Let Θ_1 denote the set of types that reach an agreement in equilibrium: $\Theta_1 \equiv \{\theta \in \Theta \mid \pi(\theta) = 1\}$.*

(a) Less-resolved Defender types reach agreement and more-resolved types deliberately choose war (i.e., $\operatorname{clos} \Theta_1 = \{\theta \in \Theta \mid \theta \leq \hat{\theta}\}$ for some $\hat{\theta}$) if

$$\frac{K_D(h', \theta') - K_D(h', \theta'')}{\theta'' - \theta'} < 1 - R(h') \tag{3}$$

for all $h' \in h(\Theta_1)$ and all $\theta', \theta'' \in \Theta$ such that $\theta' < \theta''$.

(b) Less-resolved Defender types deliberately choose war and more-resolved types reach agreement (i.e., $\operatorname{clos} \Theta_1 = \{\theta \in \Theta \mid \theta \geq \hat{\theta}\}$ for some $\hat{\theta}$) if

$$\frac{K_D(h', \theta') - K_D(h', \theta'')}{\theta'' - \theta'} > 1 - R(h') \tag{4}$$

for all $h' \in h(\Theta_1)$ and all $\theta', \theta'' \in \Theta$ such that $\theta' < \theta''$.

Proof. We prove the first claim; the proof of the second is analogous. Consider a direct mechanism that satisfies (IC) in which a less-resolved Defender type deliberately chooses war and a more-resolved type reaches agreement—i.e., $\pi(\theta') = 0$ and $\pi(\theta'') = 1$, where $\theta' < \theta''$. Incentive compatibility for θ' implies

$$\theta' \geq (1 - R(h(\theta'')))V_D(\theta'') + R(h(\theta''))\theta' - K_D(h(\theta''), \theta'),$$

while incentive compatibility for θ'' implies

$$(1 - R(h(\theta'')))V_D(\theta'') + R(h(\theta''))\theta'' - K_D(h(\theta''), \theta'') \geq \theta''.$$

Combined these imply

$$\frac{K_D(h(\theta''), \theta') - K_D(h(\theta''), \theta'')}{\theta'' - \theta'} \geq 1 - R(h(\theta'')).$$

Therefore, Equation 3 does not hold. □

C.6 Proof of Proposition 5

Proposition 5. *Suppose the linearity assumptions hold and $r - 1 < \underline{k} \leq \overline{k} < r$.*

(a) Equation 1 and Equation 3 hold.

(b) Let π^* be any absolutely continuous, weakly decreasing function such that $\pi^*(\theta) > 0$ for all $\theta \in \Theta$, and let h^* be any absolutely continuous function that satisfies

$$\frac{dh^*(\theta)}{d\theta} \geq \frac{\frac{1}{r - \underline{k}} - h^*(\theta)}{\pi^*(\theta)} \cdot \frac{d\pi^*(\theta)}{d\theta}$$

for all $\theta \in \Theta$ at which h^* and π^* are differentiable. There is an incentive compatible direct mechanism in which $\pi(\theta) = \pi^*(\theta)$ and $h(\theta) = h^*(\theta)$ for all $\theta \in \Theta$.

(c) Let π^{**} be any absolutely continuous, weakly increasing function such that $\pi^{**}(\theta) > 0$ for all $\theta \in \Theta$, and let h^{**} be any absolutely continuous function that satisfies

$$\frac{dh^{**}(\theta)}{d\theta} \geq \frac{\frac{1}{r - \overline{k}} - h^{**}(\theta)}{\pi^{**}(\theta)} \cdot \frac{d\pi^{**}(\theta)}{d\theta}$$

for all $\theta \in \Theta$ at which h^{**} and π^{**} are differentiable. There is an incentive compatible direct mechanism in which $\pi(\theta) = \pi^{**}(\theta)$ and $h(\theta) = h^{**}(\theta)$ for all $\theta \in \Theta$.

Proof. Claim (a). Under the linearity assumptions, Equation 1 is equivalent to

$$\frac{k(\theta'') - k(\theta')}{\theta'' - \theta'} < r.$$

The assumption that $k' < r$ ensures that this holds. Meanwhile, Equation 3 is equivalent to

$$\frac{k(\theta'') - k(\theta')}{\theta'' - \theta'} > r - \frac{1}{h'}.$$

Because $\max \mathcal{H} = 1$, the assumption that $k' > r - 1$ ensures that this holds for all $h' \in \mathcal{H}$.

Preliminary to claims (b) and (c). We omit the proofs that U_D must satisfy a local incentive compatibility condition given by the envelope theorem and that V_D can always be chosen to satisfy this condition given $V_D(\underline{\theta})$; these are analogous to the proofs of Lemma C.1 and Lemma C.2 above. In the general case where $\pi(\theta) \in [0, 1]$, the envelope condition is

$$\begin{aligned} U_D'(\theta) &= \left.\frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta}\right|_{\theta'=\theta} \\ &= \pi(\theta)\left[R(h(\theta)) - \frac{\partial K_D(h(\theta), \theta)}{\partial \theta}\right] + 1 - \pi(\theta). \end{aligned}$$

Under the linearity assumptions, this simplifies further to

$$U_D'(\theta) = \pi(\theta)h(\theta)\left[r - k'(\theta)\right] + 1 - \pi(\theta).$$

We now obtain the derivative of Φ_D with respect to the reported type θ', which will allow us to verify global incentive compatibility. For all $\theta, \theta' \in \Theta$, we have

$$\begin{aligned} \Phi_D(\theta' \mid \theta) &= \pi(\theta')\left[(1 - rh(\theta'))V_D(\theta') + rh(\theta')\theta - k(\theta)h(\theta')\right] + (1 - \pi(\theta'))\theta \\ &= U_D(\theta') + [1 - \pi(\theta') + r\pi(\theta')h(\theta')](\theta - \theta') - \pi(\theta')h(\theta')[k(\theta) - k(\theta')]. \end{aligned}$$

Let h and π be absolutely continuous. $\Phi_D(\cdot \mid \theta)$ is absolutely continuous and thus differentiable almost everywhere, per the same argument as in the proof of Lemma C.3 above. At each point of differentiability,

$$\begin{aligned} \frac{\partial \Phi_D(\theta' \mid \theta)}{\partial \theta'} &= 1 - \pi(\theta') + r\pi(\theta')h(\theta') - \pi(\theta')h(\theta')k'(\theta') \\ &\quad + [r(\pi h)'(\theta') - \pi'(\theta')](\theta - \theta') - [1 - \pi(\theta') + r\pi(\theta')h(\theta')] \\ &\quad - (\pi h)'(\theta')[k(\theta) - k(\theta')] + \pi(\theta')h(\theta')k'(\theta') \\ &= [r(\pi h)'(\theta') - \pi'(\theta')](\theta - \theta') - (\pi h)'(\theta')[k(\theta) - k(\theta')] \\ &= (\pi h)'(\theta')\left[r(\theta - \theta') - k(\theta) + k(\theta')\right] - \pi'(\theta')(\theta - \theta') \\ &= \left((\pi h)'(\theta')\left[r - \frac{k(\theta) - k(\theta')}{\theta - \theta'}\right] - \pi'(\theta')\right)(\theta - \theta'). \end{aligned} \tag{C.6}$$

Note that $r - \frac{k(\theta) - k(\theta')}{\theta - \theta'} \in [r - \overline{k}, r - \underline{k}] \subseteq (0, 1)$ for all distinct $\theta, \theta' \in \Theta$.

Claim (b). Suppose π is weakly decreasing and that $h'(\theta') \geq \left[\frac{1}{r - \underline{k}} - h(\theta')\right] \frac{\pi'(\theta')}{\pi(\theta')}$ for all θ' at which h and π are differentiable. Because $\pi'(\theta') \leq 0$, this implies that for all $\theta \in \Theta$,

$$(\pi h)'(\theta') = \pi'(\theta')h(\theta') + \pi(\theta')h'(\theta') \geq \frac{\pi'(\theta')}{r - \underline{k}} \geq \frac{\pi'(\theta')}{r - \frac{k(\theta)-k(\theta')}{\theta-\theta'}},$$

so the first term in parentheses in Equation C.6 is non-negative. Therefore, for all $\theta, \theta' \in \Theta$ such that $\Phi_D(\theta' \mid \theta)$ is differentiable in θ', we have that $\theta > \theta'$ implies $\frac{\partial \Phi_D(\theta'|\theta)}{\partial \theta'} \geq 0$ and $\theta < \theta'$ implies $\frac{\partial \Phi_D(\theta'|\theta)}{\partial \theta'} \leq 0$. Because $\Phi_D(\cdot \mid \theta)$ is absolutely continuous, this implies that the direct mechanism satisfies (IC).

Claim (c). Suppose π is weakly increasing and that $h'(\theta') \geq \left[\frac{1}{r-\bar{k}} - h(\theta')\right] \frac{\pi'(\theta')}{\pi(\theta')}$ for all θ' at which h and π are differentiable. Because $\pi'(\theta') \geq 0$, this implies that for all $\theta \in \Theta$,

$$(\pi h)'(\theta') = \pi'(\theta')h(\theta') + \pi(\theta')h'(\theta') \geq \frac{\pi'(\theta')}{r - \bar{k}} \geq \frac{\pi'(\theta')}{r - \frac{k(\theta)-k(\theta')}{\theta-\theta'}}.$$

The first term in parentheses in Equation C.6 is again non-negative, and the proof of incentive compatibility follows as in the last step. □

C.7 Proof of Proposition 6

Proposition 6. *In the model where D's type is prize value, if $Q_C(\beta_D) = Q_D(\beta_D) = 0$ for all β_D, then the direct mechanism is isomorphic to a contest of capabilities in the baseline framework in which Equation 1 is satisfied. If additionally $\kappa_D(h(\beta_D)) = 0$ for all β_D, then it is isomorphic to a contest of nerves.*

Proof. Take an incentive compatible mechanism that satisfies (IR-D). Define a corresponding contest of capabilities as follows:

- Map prize-value types β_D into war-payoff types $\theta(\beta_D)$ as follows:

  $$\theta(\beta_D) = \frac{-n_D}{\beta_D}.$$

  Because $-n_D < 0$, this is a strictly increasing (and thus invertible) function.

- Define the hassling cost function $K_D(h, \theta)$ as

  $$K_D(h, \theta) = \frac{\theta \kappa_D(h)}{-n_D}.$$

  Notice that this is weakly decreasing in θ for $\theta < 0$, so Equation 1 is satisfied. It is constant in θ if $\kappa_D = 0$, in which case the transformed model is a contest of nerves per our definition.

Now define a direct mechanism in the baseline framework where $\tilde{\pi}(\theta) = 1$, $\tilde{h}(\theta) = h(-\frac{n_D}{\theta})$, and $\tilde{V}_D(\theta) = S_D(-\frac{n_D}{\theta})$. Observe that $\tilde{\Phi}_D(\cdot \mid \theta(\beta_D)) \propto \Psi_D(\cdot \mid \beta_D)$ for all β_D: for any

$\theta' \in [\theta(\underline{\beta}_D), \theta(\overline{\beta}_D)]$,

$$\begin{aligned}\tilde{\Phi}_D(\theta' \mid \theta(\beta_D)) &= (1 - R(\tilde{h}(\theta')))\tilde{V}_D(\theta') + R(\tilde{h}(\theta'))\theta(\beta_D) - K_D(\tilde{h}(\theta'), \theta(\beta_D)) \\ &= (1 - R(\tilde{h}(\theta')))\tilde{V}_D(\theta') + \left[R(\tilde{h}(\theta')) + \frac{\kappa_D(\tilde{h}(\theta'))}{n_D}\right]\theta(\beta_D) \\ &= (1 - R(h(\beta'_D)))S_D(\beta'_D) + \left[R(h(\beta'_D)) + \frac{\kappa_D(h(\beta'_D))}{n_D}\right] \cdot \frac{-n_D}{\beta_D} \\ &= (1 - R(h(\beta'_D)))S_D(\beta'_D) - \frac{R(h(\beta'_D))n_D}{\beta_D} - \frac{\kappa_D(h(\beta'_D))}{\beta_D} \\ &= \frac{\Psi_D(\beta'_D \mid \beta_D)}{\beta_D}\end{aligned}$$

where $\beta'_D \equiv -\frac{n_D}{\theta'}$. Incentive compatibility of (Q_C, Q_D, h, S_D) therefore implies incentive compatibility of $(\tilde{\pi}, \tilde{h}, \tilde{V}_D)$. Additionally, because $\theta(\beta_D) < 0$ for all β_D, the latter mechanism trivially satisfies voluntary agreements because the former satisfies (IR-D). □

C.8 Proof of Proposition 7

Proposition 7. *In the model where D's type is prize value, there exist $\tilde{\beta}, \hat{\beta} \in [\underline{\beta}_D, \overline{\beta}_D]$ such that:*

(a) For all $\beta_D < \tilde{\beta}$, $Q_D(\beta_D) = 1$ and $\kappa_D(h(\beta_D)) = 0$.

(b) For all $\beta_D \in (\tilde{\beta}, \hat{\beta})$, $Q_C(\beta_D) = Q_D(\beta_D) = 0$. $h(\beta_D)$ and $S_D(\beta_D)$ are weakly increasing on this interval of types.

(c) For all $\beta_D > \hat{\beta}$, $Q_C(\beta_D) = 1$. There exists $\hat{\kappa}$ such that $\kappa_D(h(\beta_D)) < \hat{\kappa}$ for all $\beta_D < \hat{\beta}$ and $\kappa_D(h(\beta_D)) = \hat{\kappa}$ for all $\beta_D > \hat{\beta}$.

Proof. First, consider types β'_D, β''_D such that $Q_D(\beta'_D) = 1$ and $Q_C(\beta''_D) = Q_D(\beta''_D) = 0$. Incentive compatibility for each of these types implies

$$S_D(\beta''_D)\beta''_D \geq R(h(\beta''_D))n_D + \kappa_D(h(\beta''_D)) - \kappa_D(h(\beta'_D)) \geq S_D(\beta''_D)\beta'_D,$$

which in turn implies $\beta''_D > \beta'_D$. Now consider a third type β'''_D such that $Q_C(\beta'''_D) = 1$. Incentive compatibility for β''_D and β'''_D implies

$$[1 - S_D(\beta''_D)]\beta'''_D \geq \kappa_D(h(\beta'''_D)) - \kappa_D(h(\beta''_D)) - R(h(\beta''_D))n_D \geq [1 - S_D(\beta''_D)]\beta''_D,$$

which in turn implies $\beta'''_D \geq \beta''_D$. This proves that the type space can be partitioned into (potentially empty) intervals in which the lowest types of D quit, the highest types induce C to quit, and in between neither state quits.

The claim in (a) that $Q_D(\beta_D) = 1$ implies $\kappa_D(h(\beta_D)) = 0$ is immediate from (IR-D).

To prove the claim in (b) that h and S_D are weakly increasing, consider types $\beta'_D, \beta''_D \in (\tilde{\beta}, \hat{\beta})$ such that $h(\beta'_D) < h(\beta''_D)$. Incentive compatibility for both types implies

$$\begin{aligned}
&[S_D(\beta''_D) - S_D(\beta'_D)]\beta''_D \\
&\qquad \geq [R(h(\beta''_D)) - R(h(\beta'_D))]n_D + \kappa_D(h(\beta''_D)) - \kappa_D(h(\beta'_D)) \\
&\qquad\qquad \geq [S_D(\beta''_D) - S_D(\beta'_D)]\beta'_D.
\end{aligned}$$

$h(\beta'_D) < h(\beta''_D)$ implies that the middle term of the above expression is strictly positive, so the first inequality implies $S_D(\beta''_D) > S_D(\beta'_D)$. This in turn implies $\beta''_D > \beta'_D$.

Finally, to prove the claims about $\hat{\kappa}$ in (c), consider types $\beta'_D, \beta''_D > \hat{\beta}$. Incentive compatibility for these two types implies

$$\kappa_D(h(\beta''_D)) - \kappa_D(h(\beta'_D)) \geq 0 \geq \kappa_D(h(\beta'_D)) - \kappa_D(h(\beta''_D)),$$

so $\kappa_D(h(\beta'_D)) = \kappa_D(h(\beta''_D)) = \hat{\kappa}$. Additionally, incentive compatibility for any $\beta_D < \hat{\beta}$ implies

$$S_D(\beta_D)\beta_D - R(h(\beta_D))n_D - \kappa_D(h(\beta_D)) \geq \beta_D - \hat{\kappa}.$$

Because $S_D(\beta_D) < 1$, this implies $\hat{\kappa} > \kappa_D(h(\beta_D))$. □